

Poverty and transgenic crops

Africa's rejection of genetically modified food aid reflects a chasm of misunderstanding that is only exacerbated by exaggerated claims for the benefits of the technology.

At first glance, the reluctance of some nations in southern Africa to accept international donations of genetically modified (GM) maize will strike some as bizarre and irresponsible. But the drama that is now being played out across the region (see page 571) raises some serious issues to which sceptics should pay heed, before they dismiss the problem as just another example of African governance gone awry.

The first issue is the extent to which aid donors like to enjoy most of the fruits of their own benevolence. In the case of US food aid, including some of the emergency aid currently flowing into southern Africa, grants or loans are normally made available only for the procurement of grain from US farmers. That makes the decision to grant the aid more politically palatable, because it is, in effect, just a few dollars more on top of the billions already being lavished on domestic farm support.

The second issue raised by the impasse is the extent to which transgenic crops are a relevant tool in eradicating poverty and feeding the world. Supporters of transgenic agriculture are engaged in an elaborate campaign to convince the public, particularly in Europe, that this is indeed the case. But the real impact of the technology on global poverty — now and for the foreseeable future — is to increase yields in rich countries, adding to a global grain glut that depresses prices and undermines agriculture in the poorest countries.

It is only in the longer term that transgenic technology holds out promise for these countries, and this will be fulfilled only if the countries concerned can introduce the technology on their own terms. The biosafety protocol of the Convention on Biological Diversity, for example, acknowledges the need for poor nations to develop the capacity to assess new agricultural technologies for themselves, if they are to use them effectively.

South Africa is the only country in sub-Saharan Africa that has been able to plan and implement rules for the commercial growing of GM food. Other countries in the region are now being pressed to accept the technology in an emergency, effectively without informed consent. Aid agencies cannot always prevent donated grain from being sold on the black market for planting by wealthier farmers in areas not afflicted by drought, they say, so the transgenic crops will arrive in their farms by default.

Countries in the region are understandably concerned about the trade implications of this. European consumers remain wary of GM food, and tighter labelling requirements for it are in the pipeline (see *Nature* **418**, 114; 2002). Some countries are already targeting niche markets for non-GM food exports, and small African nations — whose most lucrative potential export markets are in Europe — need to keep their options open.

It is certainly to be hoped that the United States is not using the current famine threat to get its GM crops into Africa through the back door to expand the restricted export market for them.

The United States donates almost 60% of the world's food aid and, as long as much of that aid is tied to the procurement of food from US farmers, the region facing famine will probably have to accept GM food. For now, negotiations are taking place to arrange the milling of GM maize before it arrives in Zimbabwe, to prevent the possibility of replanting. Some experts even suggest that the United States could exchange its GM maize with non-GM grain from a country prepared to accept both, such as India or South Africa, so that the latter can be distributed in the famished region. Such an arrangement might sound overelaborate, but it is only a taste of what is to come if Europe and the United States continue their mutual impasse on acceptance of this technology. ■

Integrity from the top down

Research universities and other institutions are responsible for creating an environment that fosters scientific integrity.

Last month, in a rare pronouncement by the US scientific establishment on a thorny subject, the Institute of Medicine (IOM) released a report on integrity in scientific research.

The IOM notes that fully fledged cases of scientific misconduct are rare. But it sensibly calls for research institutions to take a more active role in creating an environment where misconduct will not occur.

Every scientist would agree that good science requires solid experimental design, truthful and thorough reporting of results, honest peer review, good care of living subjects, and fairness to one's colleagues and students. The IOM report looks at what can be done to encourage researchers to adhere to such nostrums throughout their careers.

It concludes that government mandates are unlikely to imbue ethics. Nor are classes in which students are herded into a lecture hall to fulfil a requirement on a checklist. Advising and mentoring are crucial but, as the report states, the quality of mentors varies immensely. So it suggests that "both the call for change and its implementation must come from research institutions".

But having accepted that, what are universities to do? The IOM suggests that they teach research ethics creatively, as an integral part of core course content. It recommends that professional bodies evaluate institutional integrity as part of the accreditation process. However, no one knows how institutional characteristics influence research integrity, and the report calls for more research on this linkage.

This assessment comes hard on the heels of two recent cases of alleged misconduct, at the University of California, Berkeley, and at Bell Labs in Murray Hill, New Jersey. Institutions would be wise to take its premise seriously. Whenever a researcher commits major fraud, he or she has probably been getting away with smaller lapses for years — fudging a control here, deleting a messy data point there.

It is in the interests of every university and laboratory to help students think through the long-term consequences of what may at first appear to be minor violations of integrity. Ultimately, such early consideration will be good for institutions, good for the careers of young researchers, and good for science itself. ■





All change
The money's in —
but does the NIH
need restructuring?
p572



Major revision
NSF instructed to set
priorities for its
main projects
p573



Viking battle
Researchers lock
horns over
ancient map
p574



Blood ties
Sheep study offers
evidence for CJD
transmission
p576

Africa hungry for conventional food as biotech row drags on

Natasha McDowell, London

Famine-relief efforts in drought-stricken southern Africa are this week trapped in the increasingly bitter and polarized global argument over the acceptability of genetically modified (GM) crops.

Most attention has focused on Zimbabwe's decision earlier this month to accept a shipment of GM maize (corn) sent as a gift by the United States only if it was first milled and therefore could not be planted. Previously, it had refused the shipment altogether.

Zambia, Namibia and Mozambique are also to varying degrees resisting the importation of transgenic crops.

But as hundreds of thousands of people face starvation in the coming months, the diverse disputes on food aid reflect a broader impasse between Europe and the United States over the perceived safety of transgenic crops. The aid is usually donated as a gift or in the form of loans or grants to purchase food from the donor country.

Some aid officials accuse international, non-governmental aid organizations of stirring up unfounded concerns in Africa about transgenic foods.

The aid organizations and African government officials counter with the argument



Relief: transgenic food has been welcomed in Malawi despite causing problems in other countries.

that the United States is using this crisis to force African countries to accept transgenic agriculture when they lack the means to independently assess the risks it may pose to the environment and to health.

Zambia is still in negotiations with the

United States about a major loan from Washington tied to the purchase of US-grown GM maize. As is common with much foreign aid, most food aid from the United States comes in the form of loans to buy food from US farmers. Mozambique has an

US calls for early data on transgenic crop safety

Erika Check, Washington

The Bush administration has released a new policy that asks companies to voluntarily report data on the safety of genetically modified (GM) crops while they are still in early-stage field trials.

The policy, released by the White House on 2 August, says that the increasing number of GM crops in field trials "could result in intermittent, low levels of biotechnology-derived genes ... occurring in commerce that have not gone through all applicable regulatory reviews".

It says that the US Food and Drug Administration (FDA) should collect information, supplied on a voluntary basis,

about the "potential toxicity and allergenicity" of proteins introduced into experimental crops. The data would be collected only if the protein had not previously been assessed in another crop.

The FDA would encourage growers "to consult with the agency about whether the presence in food/feed of such material at low and intermittent levels would raise any potential safety issues", the policy states.

Industry organizations have welcomed the voluntary requirement, saying that their members will comply, and that this will help to reassure countries that import food from the United States.

"It's good for our trading partners and

for consumers who are still gaining confidence in this new technology to know it's being looked at in very early stages," says Lisa Dry, a spokeswoman for the Biotechnology Industry Organization.

But environmental groups have attacked the policy. Matt Rand of the National Environmental Trust says that the Bush administration has now acknowledged that experimental GM crops are likely to contaminate other crops, but is not requiring companies to prove that the crops are safe.

"The international community is going to be even more sceptical of US exports than they were before," Rand says. ■

official policy of accepting only GM-free maize, as has Namibia. But Malawi has accepted free US food without raising any concerns, as have Lesotho and Swaziland.

"People here are following the global debate on GM crops and are concerned that not much is known," says Mwanya-Lewanika, a biotechnologist at the National Institute for Scientific and Industrial Research in Lusaka, Zambia. "We can't introduce GM technology without a biosafety regulatory framework in place. Until then we would prefer to buy crops from where we know they are GM-free, even if they are more expensive."

The African nations are also concerned that their future chances of exporting their own crops to Europe could be damaged if the GM grain delivered as food aid were replanted and entered the food cycle. Many European food manufacturers refuse to accept GM food, owing to consumers' dislike of the technology. Even fewer are likely to accept it if new rules come into effect that require the labelling of foodstuffs containing GM ingredients (see *Nature* **418**, 114; 2002).

"African countries now face new export hurdles because of regulatory uncertainty in Europe," says Calestous Juma, a development expert at Harvard University. "The issue is not about whether GM crops are safe or not. It is about the urgent need to agree on a predictable and non-discriminatory trading regime for GM products."

US officials claim that they could not give countries GM-free crops even if they wanted to, as US farmers do not routinely segregate GM and non-GM crops, except for the organic market.

Critics of the US aid strategy contend that it exploits the crisis by depriving the African countries of the chance to decide whether or not they want the technology. "Accepting GM technology now could stop these countries getting back on their feet in the long term," says Hannah Crabtree of the UK charity ActionAid.

Some aid officials working in Africa claim that the Zambian government is being encouraged by European aid groups to reject the US loan.

"I think it is absolutely irresponsible unless they put their money where their mouth is and come up with non-GM food," says one aid official, who asked not to be named. "I don't have the nerve, heart or soul to deny, as a precautionary principle, food to people who are hungry right here, right now. It is a debate that only America and Europe can afford because they have food."

The World Food Programme, the United Nations agency responsible for coordinating food aid, has so far received only a quarter of the US\$507 million of food aid that it has requested for the region.

Future of the NIH may lie in restructuring, committee told

Erika Check, Washington

With next year's budget topping \$27 billion, the National Institutes of Health (NIH) has every right to feel flush. But according to two of the agency's former directors, it needs to be restructured in order to spend its new-found wealth wisely.

The case for reorganization was made on 30 July, at the first meeting of a committee appointed by Congress to assess the NIH's structure. Former NIH directors Harold Varmus and Bernadine Healy told the committee, which is being run jointly by the Institute of Medicine and the National Academy of Sciences, that Congress could make the NIH more effective by reorganizing it into clusters.

The NIH was constituted as a single research institute 72 years ago. Today, it consists of 27 separate institutes and centres, each of which receives its own budget from Congress. And the NIH director has only limited direct control over the institutes. Varmus, who held the position from 1993 to 1999, has argued that this decentralized structure wastes resources and stops the director from taking decisive action, such as injecting extra resources into cutting-edge projects.

Varmus and Healy told the meeting that reorganization would allow the NIH to respond to new research needs and eliminate administrative duplication. Varmus suggested that legislators start by creating a National Institute of Brain Disorders, which would incorporate six current institutes, including the National Institute of Mental Health and the National Institute of Neurological Disorders and Stroke. "We have a pretty good opportunity to do an experiment here," he said, arguing that if the new cluster was successful, others could follow.

The House of Representatives subcommittee that funds the NIH requested the study in its 2001 appropriations bill. And although any threat to the institutes' cherished auton-



Grand design: originally a single institution, the NIH now consists of 27 institutes and centres.

omy is likely to be opposed by individual congressmen, advocacy groups and even researchers, the committee was told by congressional staffers that its recommendations could help to overcome such resistance.

"You are protected from certain pressures that are on us," said David Bowen, a staff member of the Senate committee that oversees the NIH. "It will be far easier for you to come up with a proposal than it would be to originate it on Capitol Hill," he said.

In addition to looking at the NIH's overall structure, the panel, which is chaired by Harold Shapiro, former president of Princeton University, is being asked to consider a few specific questions. Cheryl Jaeger, a member of the House Energy and Commerce Committee in the House of Representatives, asked the panel to consider whether the Institute on Drug Abuse and the Institute on Alcohol Abuse and Alcoholism could be merged. And Bowen asked the panel to contemplate what should happen to the National Human Genome Research Institute as the Human Genome Project nears completion.

Former congressman John Porter, who used to chair the subcommittee that requested the report, warned panel members to concentrate on changes that are politically realistic. He said a proposal for new clusters is more likely to be implemented if each institute keeps some measure of autonomy.

And current NIH director Elias Zerhouni cautioned the panel to think seriously about whether reorganizing the agency would do any good, when it already functions relatively well.

Others said that the panel, which will report in 2003, must rapidly enlist the support of the health and science advocacy groups that have a stake in the NIH. "They should consult widely, because if the committee does see fit to recommend major changes, then we can be certain that there will be major resistance," said Leon Rosenberg, a molecular biologist at Princeton University who headed a previous panel on how the NIH should set research priorities.



Fresh focus: Bernadine Healy (left) and Harold Varmus advocate reorganizing the NIH.

Big projects cited as threat to NSF's integrity

Geoff Brumfiel, Washington

The integrity of the US National Science Foundation (NSF) could be called into question because of its backlog of competing major projects, observers claim.

The warning came at a hearing on 2 August of a National Academy of Sciences panel set up to study how the NSF selects such projects. Congressional staff warned the panel that the number of approved but so far unfunded projects was encouraging Congress to choose those that the NSF should fund. This could cost the agency its cherished reputation for scientific objectivity, they said, and perhaps lead it into territory where political factors, rather than scientific merit, would determine project funding.

But NSF director Rita Colwell denies that the problem exists, and says that the agency's selection process is in good shape.

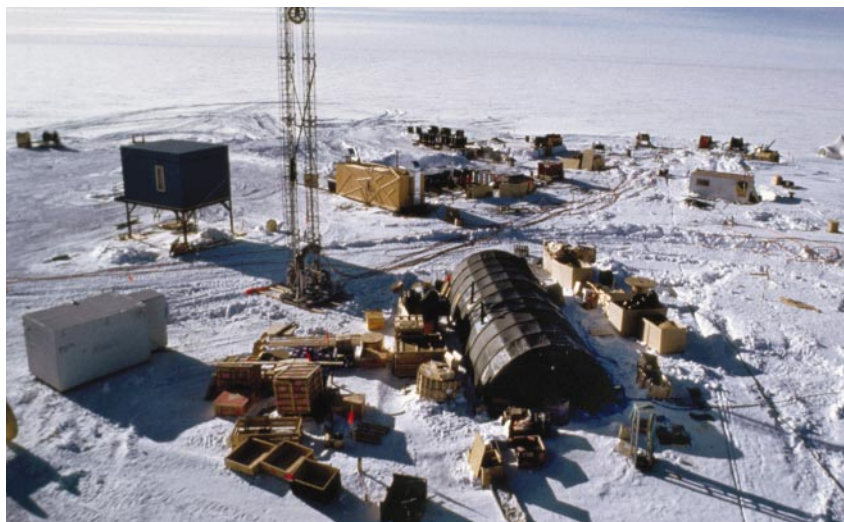
In the past few years, the National Science Board, which governs the NSF, has approved several multimillion-dollar projects that the agency has not felt able to fund for construction. At least six projects are currently in this backlog — and researchers involved in them are increasingly bypassing the NSF and going straight to Congress to seek funding, staff told the meeting.

One such project is IceCube, a neutrino detector to be built in the Antarctic. This was approved by the science board in 2001, but the NSF has yet to request funds for it. Francis Halzen of the University of Wisconsin at Madison, who directs the project, met with the staff of the House of Representatives appropriations subcommittee that funds the NSF — “at their request”, he says — to discuss IceCube. As a result, he secured \$15 million to begin work on the project this year.

Another project, the High-performance Instrumented Airborne Platform for Environmental Research (HIAPER), which will study the troposphere using an aircraft, was approved by the science board in 1998, but is yet to appear in an NSF budget request. Congress, however, added \$35 million for the project to this year's NSF budget.

Congressional staff also reported strong pressure to squeeze funding for a proposed \$280-million underground physics laboratory at the Homestake mine near Lead, South Dakota, into the NSF's budget. The project has yet to undergo scientific review, but it is strongly backed by both main political parties, who are facing off in a fiercely competitive Senate election in South Dakota this November. One congressional staff member, claimed that pressure to begin the project is threatening to bypass the entire approval process of the National Science Board.

“The volume of people coming to Capitol Hill from the science community is increasing,” Joel Widder, a staff member on the Sen-



The IceCube project has jumped its place in the funding queue, thanks to congressional intervention.

ate subcommittee responsible for NSF's budget, told the panel. Cheh Kim, another staff member, asked the panel — which was convened at the request of a group of senators who say they champion the NSF's independence — to help to establish criteria for prioritizing NSF projects. Meantime, Senate appropriators have slashed next year's funding for major

NSF projects (see *Nature* **418**, 472; 2002).

But Colwell contends that the problem of congressional interference with NSF project decisions is under control. “I think the process is working very well and has been for 50 years,” she says. She doubts whether the academy's review will lead to any drastic changes in the process. ■

Test-ban treaty ‘scientifically sound’

Geoff Brumfiel, Washington

The Comprehensive Nuclear-Test-Ban Treaty (CTBT), rejected by the Senate nearly three years ago, has been pronounced technically sound by a panel of leading US scientists.

The treaty, which aims to prohibit nuclear testing worldwide, was signed by then president Bill Clinton in 1996, but the Senate voted against its ratification in October 1999.

Opponents of the CTBT — including leading members of the Bush administration — have argued that the treaty's terms would not allow the United States to detect testing abroad, and that adherence to it would threaten the country's own nuclear capability. Some want the administration to withdraw its signature from the unratified treaty.

But in a study released on 31 July, the National Academy of Sciences takes issue with these arguments. John Holdren, director of a science, technology and public-policy programme at Harvard University, and the chairman of the committee that drafted the report, says that using the monitoring system stipulated by the treaty, the United States would be able to detect tests as low as 1 or 2 kilotons virtually

anywhere on Earth or in space. Attempts to mask a blast would fail unless the country had extensive previous testing experience. And such countries, including Russia and China, would have little to gain from low-yield, clandestine testing, he claims.

The study also finds that the United States could maintain its own stockpile of nuclear weapons indefinitely without further testing.

“This is the most authoritative and detailed assessment of the test-ban treaty to date,” says Daryl Kimball, executive director of the Arms Control Association, which advocates ratification of the test ban. “It should give the Bush administration reason to reconsider its fundamental points of contention on the CTBT.”

But Ivan Eland, director of defence policy studies at the Cato Institute, a libertarian think-tank, doubts whether it will have much impact on administration policy. “To me, it doesn't make sense to sign up for a treaty if you might have to test new weapons,” he says.

The Bush administration opposes ratification of the treaty, but it continues to observe the US moratorium on testing that was begun in 1992. ■

♦ www.nap.edu/books/0309085063/html

Ink analysis raises storm over Viking map

John Whitfield and Rex Dalton

Scientists are locking horns over a Viking map of America. One group claims to have evidence that the document is a fake, but another says that the parchment on which it is drawn predates Christopher Columbus's 1492 discovery of the New World.

The Vinland Map's ink dates to the twentieth century, argues one team — adding that the parchment's age is irrelevant. But other researchers have roundly criticized the ink analysis, saying that the carbon dating of the parchment nudges researchers closer to the conclusion that the map is genuine.

If it is, then the parchment, which depicts Leif Ericsson's journey to what is now Canada in the tenth century AD, is the oldest map of America. It came to light in 1957, when Yale University paid \$1 million for it. The idea that the Vikings made it to America before Columbus was new and controversial at the time, although archaeological findings have since confirmed it.

But some have always doubted the map's authenticity, and in 1974 crystals taken from the map were found to contain anatase, a form of titanium dioxide used in white paint which has only been manufactured since 1923.

Now chemists Robin Clark and Katherine Brown of University College London have studied the map's ink with a technique called laser Raman microscopy, which uses laser scattering to reveal chemical composition. The map's lines are made of two different substances, they report¹. Their black centres are carbon; their yellowish edges are composed of anatase. Clark says that anatase shouldn't be present in a map from the fifteenth century. He thinks that a forger overlaid a yellow line



Fact or fake? The Vinland Map's age is still disputed after conflicting analyses of its ink and parchment.

with a black one to create an aged appearance.

But other experts have blasted these conclusions. They are "singularly unconvincing," says physicist Thomas Cahill of the University of California, Davis. "There's no example of any forger trying to do double inking."

Jill Pasteris of Washington University in St Louis, Missouri, a leading authority on laser Raman microscopy, says that the London team has overstated its results, and that the researchers did not even use a microscope to scan the parchment properly for anatase.

"The presence of anatase is by no means conclusive proof that the map is a forgery," says Jacqueline Olin, a retired Smithsonian Institution chemist who has studied the map for 30 years. More analysis is needed to answer

the anatase question decisively, she says.

Olin is part of a team led by Garman Harbottle, a specialist in carbon dating at Brookhaven National Laboratory in Upton, New York, that has dated the map's parchment to between 1423 and 1445 (ref. 2).

"It doesn't prove the map's authenticity, but it's another piece of evidence to be put in the scales," says Harbottle. "It sets the bar higher for those who want to show that the map's a fake." Carbon-dating the ink might settle the matter, Harbottle suggests. But current carbon-dating techniques require more material than the ink could provide. ■

1. Brown, K. L. & Clark, R. J. H. *Anal. Chem.* **74**, 3658–3661 (2002).
2. Donahue, D. J., Olin, J. S. & Harbottle, G. *Radiocarbon* **44**, 45–52 (2002).

Swiss scientists fear for top ranking in basic research

Oliver Schmidt, Munich

With generous science funding and world-class facilities, Switzerland is often seen as a land of milk and honey for researchers. But paradise is in trouble, say scientists working there. Over half of the scientists questioned by the Swiss National Science Foundation (SNSF), the main Swiss research funding agency, fear that the country is losing ground in international competition.

"Our leading position is endangered," agrees Konrad Basler, a molecular biologist who is based at the University of Zurich.

The SNSF's findings come in a report, released on 30 July, detailing a survey of 4,000 researchers — around half of Swiss science's workforce. Fears for the status of Swiss research also seem to be backed up by citation statistics for Swiss research papers.

Markus von Ins of the Centre for Science and Technology Studies in Bern says Swiss dominance in the National Science Indicator (NSI) rankings, which measure the average scientific impact of papers produced by a particular country, is waning. Switzerland topped the NSI table from 1981 to 2001, but although its score has improved little since 1981, that of the United States, in second place, has increased by 6%, with Britain and the Netherlands also closing the gap.

Swiss scientists say that the government's short-term economic strategy is partly to blame. Annual funding for Swiss basic research rose modestly from 215 million Swiss francs (US\$146 billion) to 248 million Swiss francs between 1991 and 2001, while funding for applied research tripled to 321 million Swiss francs. "The priorities have

clearly switched from fundamental to applied research," says Adriano Aguzzi, a neuropathologist at the University of Zurich.

The SNSF, which already spends most of its money on basic research, has called for a billion Swiss francs in extra funding between 2003 and 2007. The Swiss Science and Technology Council, which advises the government on science-policy issues, has promised to support the SNSF's plea. "We recommend doubling the budget," says Gottfried Schatz, the council's president.

George Thomas, a molecular biologist at the Friedrich Miescher Institute in Basel, says that Switzerland is still an attractive place for research. "Switzerland is but a small country, and the SNSF is doing a good job," he says. "But with massive US investment in the life sciences, it's really difficult to keep on track." ■

Deer meat suspected of causing hunters' brain-disease deaths

Atlanta Epidemiologists are investigating whether the deaths of three hunters from brain diseases are linked to their possible consumption of meat infected with chronic wasting disease (CWD).

The US Centers for Disease Control and Prevention in Atlanta, Georgia, last week sent a team of researchers to Wisconsin to help officials study the deaths. The hunters often participated in wild-game feasts at a Wisconsin lodge during the early 1980s, and may have consumed elk meat from Colorado, Wisconsin state epidemiologist Jeff Davis told *Nature*. CWD, a transmissible spongiform encephalopathy similar to mad-cow disease and its human form, Creutzfeldt–Jakob disease (CJD), has been endemic in Colorado since the 1960s.

Two of the hunters died in 1993, one from CJD and the other from a more common neurological ailment, Pick's disease. The third died from CJD in 1999. There is no confirmed link between CJD and the consumption of deer or elk infected with CWD. But beef from animals with mad-cow disease is suspected of causing variant CJD, a form of CJD that has afflicted dozens of people in Britain.



Colorado elk populations are infected with chronic wasting disease, which is similar to CJD.

UCL's provost resigns as financial crisis looms

London Chris Llewellyn Smith has resigned as provost of University College London (UCL), amid reports that senior academics have expressed concern about his leadership abilities.

A physicist and former head of CERN, the European particle-physics laboratory near Geneva, Llewellyn Smith is believed to have offered to resign after the university's ruling council received a letter from dozens of senior staff expressing no confidence in him. The university announced his resignation "with regret" on 25 July, and said that the college's former provost Derek Roberts will take over for a year.

The departure is being blamed on staff anger at the college's financial problems — it is forecast to lose some £8 million (US\$12.5

million) this year — and controversial restructuring plans, including plans to merge some departments. Llewellyn Smith's position was further weakened when the college failed to perform as well as hoped in the recent research assessment exercise.

A UCL spokesperson said that Llewellyn Smith is leaving to pursue other interests, initially in the physics department at the University of Oxford.

Cell restrictions stem Framework dispute

Munich A proposal from Denmark to restrict European Union (EU) funding of stem-cell research is set to resolve the disputes that have dogged the European Commission's sixth Framework programme.

Up to 3% of the 17.5-billion-euro (US\$17.2-billion) Framework programme, which will begin in November, could originally have been spent on projects involving human embryonic stem cells. But Germany, Austria and Italy objected to such research on ethical grounds, and at one point the dispute threatened to derail the entire programme (see *Nature* 417, 680; 2002).

Denmark, which currently holds the EU presidency, proposed last week that funding for research on human embryonic stem cells should be restricted to cell lines that already exist in cell banks or in culture. The suggestion has been met with informal agreement from all the EU's member states.

The restriction will be in place until at least the end of 2003, during which time the European Commission will compile a report on the legal and scientific developments in the field. Member states will then try again to find a common position.

Dinosaur theft proves costly for palaeo-plunderer

San Diego US federal authorities have won a conviction for the theft of an *Allosaurus* skeleton from government land in Utah — but have not yet recovered the dinosaur, which was sold to a Japanese company.

Barry James — a fossil dealer in Sunbury, Pennsylvania — pleaded guilty to charges of theft by receiving stolen property, on 23 July in a state court in Salt Lake City, and agreed to pay the government \$50,000 in compensation. The conviction signals the hard line being taken by US authorities to stop the theft of palaeontological specimens from public land.

James paid two men to dig up the skeleton from federal land in Utah more than a decade ago. He sold it in 1992 to Hayashibara of Okayama, which carries out research on pharmaceuticals and food products.

Hayashibara officials were unaware that the fossil was stolen and had hoped to exhibit it. The company is now considering returning the fossil to the United States.



Research suggests there is "an appreciable risk" that vCJD can be spread in blood transfusions.

Sheep study highlights CJD risk from blood

London Researchers have produced the strongest evidence so far that prion diseases, including variant Creutzfeldt–Jakob disease (vCJD), can be transmitted through blood.

Tests at the Institute of Animal Health's neuropathogenesis unit in Edinburgh found that, of 24 sheep given blood from sheep infected with bovine spongiform encephalopathy (BSE), two developed the disease themselves (*Journal of General Virology*; doi: 10.1099/vir.0.18580-0). The same group showed transmission to a single animal two years ago (F. Houston, J. D. Foster, A. Chong, N. Hunter, & C. J. Bostock *Lancet* 356, 999–1000; 2000), but others questioned the result because only one animal was involved.

The new result shows there is "an appreciable risk" of humans catching vCJD from blood transfusions, the team says. The British government has responded to the study by considering new safety measures, such as excluding people who have received blood transfusions from giving blood.

Blooming algae pose health risk in the Baltic Sea

Stockholm Swedish holiday-makers flock to the islands and southern coastline of the Baltic Sea every summer. Now they may want to rethink their choice of destination, after mild weather triggered a surge in toxic algae.

Nodularia spumigena can cause liver damage and is potentially lethal to small children. Blooms of the algae are often seen in the Baltic Sea in summer, but recent high temperatures and low winds have allowed the algae to thrive, forcing the Swedish authorities to ban swimming in some areas.

The algae feed on nutrients found in sewage from Estonia and Russia. Pentti Kangas, a marine biologist at the Finnish Environment Institute in Helsinki, says that sewage from the St Petersburg region is washed untreated into the Baltic Sea.

♦ www.helcom.fi/environment/algalblooms.html

It's the ecology, stupid!

Can studies of environmental influences on developing organisms provide the key to understanding the evolution of species and populations? A growing number of researchers think so. Jennie Dusheck reports.

We've seen evo-devo, now prepare for the next wave: eco-devo. Evolutionary developmental biologists have already made great strides in understanding how genes that direct the formation of body plans have influenced the evolution of major animal groups¹. But organisms evolve in the context of their physical environments and of the other living things with which they interact. So biologists are now adding ecology to evo-devo's blend of evolutionary biology, developmental biology and genetics.

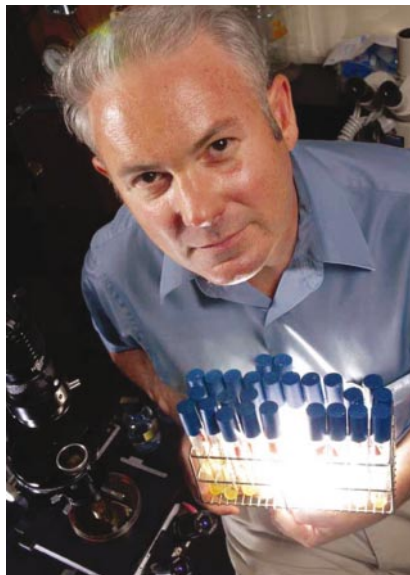
"Eco-devo is a part of evo-devo that hasn't received its due," says Scott Gilbert of Swarthmore College in Pennsylvania. But since he coined the term 'ecological developmental biology' in a review article² last year, the profile of this subdiscipline has been rising fast. "There has been a coalescing around the idea that we are going to integrate ecological context and development in a new way," says Sahotra Sarkar, a philosopher and historian of science at the University of Texas at Austin who studies trends in developmental biology.

Jessica Bolker, an evolutionary developmental biologist at the University of New Hampshire in Durham, attributes much of the upsurge in interest to Gilbert. "It's all Scott's fault," she jokes. "He's really excited about this, wants to push it, and is in a position to do so. He knows everybody."

A pivotal development came in January with a symposium on eco-devo, organized by Bolker and Gilbert, at the annual meeting of the Society for Integrative and Comparative Biology in Anaheim, California. Gilbert likens this gathering to the 1981 Dahlem Workshop on Evolution and Development in Berlin, which established baseline questions for the field of evo-devo. So what, in the light of the January symposium, is eco-devo all about?

Until now, explains Gilbert, evo-devo has focused primarily on mechanisms by which major taxonomic groups, such as phyla and classes, come into existence. But by considering development as a weaving together of genetic and environmental information to form a functioning organism, eco-devo will provide the detail — explaining more about evolution at the levels of species and subpopulations.

A simplifying assumption of standard evolutionary theory is that genetic differ-



Champions of the cause: Scott Gilbert (left), who coined the term 'eco-devo', and Jessica Bolker, the idea's other strongest proponent, kickstarted the discipline at a landmark symposium in January.

ences determine the relative success or failure of organisms in a given environment. But in practice, many organisms can respond with considerable flexibility to a changing environment. Indeed, an overriding theme of eco-devo is phenotypic plasticity — the ability of a single set of genes to generate a range of characteristics, or phenotypes, depending on the environment in which the developing organism finds itself.

Planting ideas

To plant biologists, this is old news — amateur horticulturalists know that a cutting grown in moist conditions can look very different from a genetically identical cutting grown in dry soil. So the organizers of the January symposium made sure that plant biologists were invited. "Plant people are the ones who really do eco-devo," says Bolker.

Sonia Sultan of the Wesleyan University in Middletown, Connecticut, is one such researcher. She is interested in discovering, for example, whether generalist plants, which seem to survive well in a wide variety of environments, show greater plasticity than more specialized ones. She wants to know how plasticity relates to the diversification of closely related species, and whether

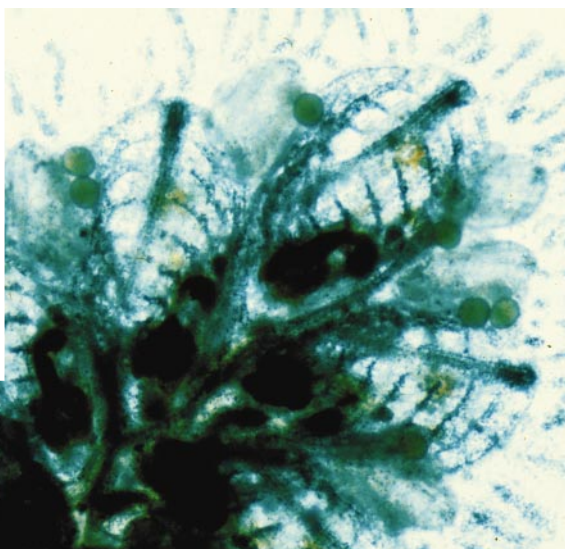
plasticity can help to explain why some types of plant are likely to become weeds while others slide towards extinction.

At the Anaheim symposium, Sultan described her studies of four closely related species of buckwheat. The four differed in the extent and form of plasticity for a variety of traits, including root length and shape, the rate of photosynthesis and leaf size³. For example, the generalist species *Polygonum persicaria* reproduced more vigorously in resource-rich environments, and doubled its amount of leaf tissue in low light conditions. In contrast, *Polygonum hydropiper*, which grows only in sunny habitats with rich, wet soil, showed far less plasticity in these traits.

Nineteenth-century developmental biologists were aware that animals can show similar plasticity. But since then, the field has lost this perspective, partly through focusing on animals that share traits, such as rapid development, that tend to minimize the effects of the environment. So the view emerged that genes 'programme' development. "Most of our models are small and fast and hard-wired, and so we think of development as being hard-wired," says Bolker. "But mice are not average mammals, *Drosophila* are not average insects, and *Xenopus* are weird frogs."

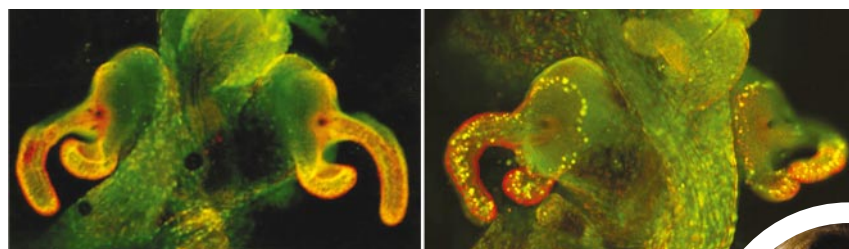


J. STEWART-SAVAGE



John Stewart-Savage (inset, right) has shown how nutrients affect reproduction in colonial sea squirts.

J. FOSTER

The bacterium *Vibrio fischeri* triggers maturation of the Hawaiian squid's light organ (above, right). Margaret McFall-Ngai has shown that cheating bugs can't do the same.

Some biologists are now staking out the territory of zoological eco-devo by studying different species. Developmental biologist John Stewart-Savage of the University of New Orleans, for instance, has teamed up with evolutionary ecologist Philip Yund of the University of Maine's Darling Marine Center in Walpole to study the heritability of reproductive effort in the colonial sea squirt *Botryllus schlosseri*. *B. schlosseri* is a cyclical hermaphrodite, each individual within the colony alternating between making eggs and producing sperm.

Stewart-Savage and Yund took genetically identical sea squirts and placed them 10 kilometres apart in the estuary of Maine's Damariscotta River. At the upstream site — a warmer, more nutrient-rich environment — individuals produced more eggs, whereas at the cooler, relatively nutrient-poor downstream site, individuals had larger testes⁵. This makes evolutionary sense: because eggs are more expensive to make, sperm production is favoured where nutrients are in short supply.

What was most striking, however, was that different genetic clones showed different degrees of plasticity. "There's a genetic basis for plasticity in reproductive effort," says Stewart-Savage. And that means that highly variable environments can select for

greater plasticity, whereas those that are more stable select for low plasticity.

Lighting the way

Other organisms can be just as powerful as the physical environment in influencing development. One of the most dramatic examples is found in the Hawaiian squid *Euprymna scolopes*. In young squid, the light-emitting bacterium *Vibrio fischeri* — a relative of the microbe that causes cholera — guides the normal development of organs that house the bacteria and illuminate the lower surface of the squids' bodies. This prevents the animals from showing up as dark shapes when viewed from below by predators.

Once the bacteria infect immature light organs in juvenile squid, the organs' morphology develops over the ensuing four days through a process of bacteria-induced programmed cell death and cell swelling. But if young *E. scolopes* are never exposed to *V. fischeri*, these changes do not occur⁶. Researchers led by Margaret McFall-Ngai of the University of Hawaii have shown that the interaction is resistant to cheating by other bacteria looking for a comfortable home. Her team created two non-light-emitting *V. fischeri* mutants; neither was able to colonize the light organ⁷.



Bacteria also influence normal development in vertebrates. For example, DNA-microarray comparisons of the activity of 25,000 genes from otherwise germ-free mice with and without the gut-living bacterium *Bacteroides thetaiotaomicron* have revealed marked changes in gene expression in the host gut when the bacterium is present⁸. Some of the 100-plus genes involved are known to influence such tasks as nutrient absorption, blood-vessel formation and maturation of the intestine. This may help to explain why germ-free mice need around 30% more food than mice raised under normal conditions — particularly given that *B. thetaiotaomicron* is just one of hundreds of species that make up a healthy gut microbial flora.

Other living influences on development are rather larger. The presence of a water-borne chemical released by predatory dragonfly larvae, for instance, induces dramatic changes in the development of wood frog (*Rana sylvatica*) tadpoles. When tadpoles grow in water that has housed the predators, they develop bigger tails that allow for faster swimming and sharper turning⁹.

Bolker says that the blossoming of evo-devo has awakened biologists to the importance of understanding developmental processes. "More people than just developmental biologists now know that development has implications for other fields," she says. And she suspects that increasing awareness of molecules in the environment that affect animals' development — chemicals that mimic the sex hormone oestrogen, for example — should provide impetus to the newer subdiscipline of eco-devo.

Gilbert hopes to answer a range of fundamental questions for a multitude of different species. "Do organisms vary in how hard-wired they are? What difference does that make? What are the genes that control plasticity? Can you select for plasticity?"

If the new practitioners of eco-devo can rise to the challenge of answering these questions, they could close the loop between ecology and evolution on the one hand, and genetics, cell biology and developmental biology on the other. "This is about how organisms do what ecologists know that they do," says Bolker. ■

Jennie Dusheck is a writer in Santa Cruz, California.

1. Arthur, W. *Nature* **415**, 757–764 (2002).
2. Gilbert, S. *Dev. Biol.* **233**, 1–12 (2001).
3. Sultan, S. E. *Ecology* **82**, 328–343 (2001).
4. Bolker, J. *BioEssays* **17**, 451–455 (1995).
5. Stewart-Savage, J., Sires, A. & Yund, P. O. in *Biology of Ascidiaceans* (eds Sawada, H., Yokosama, H. & Lambert, C. C.) 311–314 (Springer, Tokyo, 2001).
6. McFall-Ngai, M. *J. Dev. Biol.* **242**, 1–14 (2002).
7. Visick, K. L., Foster, J., Doi, J., McFall-Ngai, M. & Ruby, E. G. *J. Bacteriol.* **182**, 4578–4586 (2000).
8. Hooper, L. V. et al. *Science* **291**, 881–884 (2001).
9. Van Buskirk, J. & Relyea, R. A. *Biol. J. Linn. Soc.* **65**, 301–328 (1998).

Who's been looking at your data?

You thought that all those e-mails, data and grant proposals on your computer were for your eyes only? Think again, says Declan Butler. Someone could be snooping on your every keystroke.

For a few days in February 2000, some of the Internet's biggest sites were toppling like pins in a bowling alley. Yahoo!, Amazon, eBay, CNN.com and others were all hit by 'distributed denial-of-service attacks' launched from computers including machines at Stanford University and the University of California, Santa Barbara. The commercial sites' servers were so busy dealing with electronic junk fired at them by the academic computers that no one could gain access. It was literally child's play: the universities' computers had been hijacked by a 15-year-old Canadian known as Mafiaboy.

This breach of academic computer security was far from an isolated event. Twice in as many weeks last year, hackers broke into computers at Indiana University, accessing names, social-security numbers and addresses of thousands of students, and using the university's servers to store music and other files, and to run chat rooms. Embarrassingly, the security hole through which the intruders gained access was one that the university's computer support service had circulated warnings about only months previously. And late last month came the revelation that officials at Princeton University in New Jersey were under FBI investigation for gaining repeated unauthorized access to the online admissions system at its rival Yale University in New Haven, Connecticut.

These incidents are just the tip of the iceberg — most aren't reported because institutions don't want negative publicity. Universities and research labs are home to some of the world's most sophisticated computer systems. But the standards of computer security that prevail across large parts of academia would give a security professional sleepless nights, says Michael McRobbie, Indiana University's vice-president for information technology (IT). Computer-security offices are often "nonexistent, organizationally buried or understaffed", he says. As a result, anarchy often reigns, with staff plugging machines in and out of networks without anyone knowing, and inadequate checks on whether the people accessing data are who they say they are. Often, there are no central records of what sensitive data are being stored, and where.

Under attack

If you still find the topic a crushing bore, here is at least one reason to sit up and pay attention — the number of attacks is mushrooming. In the early 1990s, says one official at the Massachusetts Institute of Technology (MIT), a major research university might expect two or three computer-security incidents per year. Today, the number of breaches runs in the thousands. "These events, once an amusing technical curiosity,



Two faces of hacking: the annual DefCon gathering (above) highlights security flaws, whereas Mafiaboy (left) exploited them to malicious ends.

have become a significant organizational concern," says the MIT source.

In part, this is because you no longer have to be a technical wizard to break into a computer network — automated software, free for download from hacker sites, can do the job for you. 'Distributed port scans', for example, randomly crawl the Internet and probe the entry ports into networks, see what software is running, look for vulnerabilities and automatically break in. On average, each of Indiana University's 55,000 Internet-connected computers is probed at least once a day, says Mark Bruhn, the university's IT policy officer.

When computer-savvy researchers find



On guard: university computer systems need almost constant monitoring to prevent hacker activity.

that a breach has occurred, correcting the problem is no trivial matter. "I find security considerations to be a significant drag on scientific productivity," says Mark Gerstein, a bioinformatician at Yale University. "The few incidents we've had here where computers were cracked wasted a vast amount of time." Even if data remain intact, hacked computers often need to be reinstalled to ensure that the intruders have not left a secret 'back door', giving them future access.

Spies in the wires

Many of the breaches are the work of teenage hackers who want to impress their peers, but others have a more sinister edge. It is perhaps unlikely that your fiercest academic rival is surreptitiously hacking into your unpublished data, but anyone involved in a commercially sensitive project would be foolish to discount the threat of cyber-espionage. Already, some biotech firms are wary of using public databases for fear that competitors will capture their search terms and gain insights into their work. And in May this year, police in Japan arrested three staff at NEC Toshiba Space Systems in Yokohama on charges of hacking computers at the National Space Development Agency to steal designs of a satellite antenna from rival company Mitsubishi Electric.

The most serious threat of all may come from cyber-terrorists. Security experts are seriously worried that rogue states or terrorist groups could launch Mafiaboy-style denial-of-service and other attacks, but on a much larger scale, targeting infrastructures such as electricity grids and communication systems, and disabling vast swathes of the world's computers.

It's against this background that academic leaders are waking up to the need to revamp their security procedures. In a survey last year, security did not figure among the top ten IT issues ranked by the 1,800 member institutions of EDUCAUSE, which promotes the use of IT in US higher education. In the 2002 survey, security rose to number five in the rankings — and number two for medium-to-large organizations.

Institutions that aren't taking the threat seriously may soon find that external pressures force them to do so. Lawyers are pressing for research laboratories and universities to be held liable if they neglect to secure their systems and so allow their computers to be used to launch denial-of-service attacks or give hackers access to confidential data. The likes of Mafiaboy typically have few assets, so the "deep pockets of universities are an attractive target", observes Frank Vinik, risk manager with United Educators Insurance of Chevy Chase, Maryland, the main insurer for the US academic sector.

Alan Paller, director of research at SANS, the System Administration, Networking and Security Institute in Bethesda, Maryland, which brings together more than 156,000 computer professionals, suggests that government research funding should be made conditional on labs meeting minimal security requirements. And Richard Clarke, President George Bush's computer-security adviser, will next month issue a national cyberspace protection plan that is expected to include specific guidelines for government agencies. "Every American relies upon cyberspace and every American has to do something to secure their part of cyberspace," Clarke has said.

But securing academic networks is inherently difficult. A campus will typically have tens of thousands of machines hooked up to the Internet at any one time, and the corporate approach of applying rigid central control and putting everything bar the servers that run public websites behind security firewalls isn't appropriate. "University networks are by design and of necessity open," observes Gregory Jackson, vice-president and chief information officer of the University of Chicago.

What's more, many computers used for research are uniquely configured, for example for data gathering, and often are only fully understood by the scientists running the project. "Often, the system administrator knows not much more about a specific computer than where to turn it on or off," says Gerstein. Individual labs usually cannot afford to hire their own computer-security specialist, resulting in a risky do-it-yourself

approach. Stephen Nesbitt, director of operations at NASA's Computer Crimes Division, which has an enviable record in bringing malicious hackers to justice, observes that security will often come second to pressing deadlines to finish a research paper or to complete an experiment.

Despite these disadvantages, the prospects aren't entirely gloomy. Academia boasts some of the world's foremost experts in computer security. And staff at the CERT Coordination Center at Carnegie Mellon University in Pittsburgh, the main clearing-house for information about security threats, are on hand to provide technical assistance and to coordinate responses to attacks.

On the defensive

So what can researchers and their host institutions do to improve the situation? Academic networks are often stuffed with sensitive information such as research data, grant proposals, and medical and student records. The first step is to ring-fence off the most critical parts using firewalls designed to block all but authorized traffic.

Intrusions are inevitable, so both in front of the firewalls and behind them, systems should have encrypted communications — much like those you use when banking online — to protect passwords and data from prying eyes. And every single machine should be protected, or 'patched', against newly identified security flaws, and have its antivirus software up to date.

Gerstein's lab has adopted this layered approach to security. He has separated its network into private and public spaces, created secure 'private' backups of the data on servers running public websites, implemented encryption, and keeps up to date with the latest security patches. "This consumes a serious amount of time and hinders implementation of useful bioinformatics services," Gerstein admits. But given the importance of protecting data, and the hassle of restoring systems after a security breach, he argues that it is time well spent.

The adoption of encryption remains patchy, however. The University of Chicago



has secured half of its key IT services this way and expects to complete the process by autumn. But most universities have been slow to take up the technology.

Nevertheless, a few institutions stand out as models of good practice. MIT's IT staff probe its networks to spot machines that are insecure, and have the power to force those that don't make the grade off the network immediately—a policy that most institutions have been reluctant to implement. As MIT doesn't rely heavily on firewalls, this tough enforcement obliges users to install software patches to protect against new threats.

Protect and survive

This policy paid off in July last year, when the Code Red virus spread across the world's computer networks. "It was a non-event on the MIT campus," says one university IT official. Code Red exploited a security flaw in Microsoft's IIS web-server software, copying itself to randomly chosen Internet addresses, defacing websites with the message "Hacked by Chinese", and attempting a denial-of-service assault on a White House website. Variants of Code Red infected hundreds of thousands of machines. In disrupting the performance of infected systems and requiring them to be cleaned up, the viruses caused up to US\$2 billion of damage. To William Wulf, president of the US National Academy of Engineering, this experience illustrates the "Magenot line syndrome"—a false sense of security that can be engendered by firewalls.

MIT has also created a cross-institute team comprised of information-security staff, students, faculty members and representatives of its big independent labs, such as the Laboratory for Computer Science, the Artificial Intelligence Laboratory, the Media Lab, the Research Laboratory of Electronics and the Whitehead Institute for Biomedical Research. "This broad-based approach recognizes that the information-technology experts cannot handle security alone," says Vinik.

Broad-based also means thinking beyond the campus perimeter. Even if lab computers are secure, networks may be vulnerable as a result of researchers working from home, or at conferences, on computers that have been

compromised so that an intruder can capture every keystroke. This risk was famously demonstrated in October 2000 by a hacker who managed to access Microsoft's most heavily guarded asset—the source code of its Windows operating system—through an employee logging on to the company's secure system from an unprotected

home computer that had been compromised.

Mention Microsoft to any security expert, and they will point out that the company's big-selling, interconnected products are the most common point of illicit entry into a system. "The biggest and most widespread security threats have come from two or three Microsoft products, primarily Outlook and the Microsoft IIS web server, but also Internet Explorer," says John Franks, a mathematician and computer-security expert at Northwestern University in Evanston, Illinois.

Franks, for one, favours the use of products such as the Linux operating system, in which the source code is open for anyone to scrutinize, increasing the chance that security vulnerabilities will be spotted and made public. Red Hat, the major Linux distributor, has configured its software so that, when a new security patch is released, it can be automatically installed on Internet-connected machines running the operating system.

Microsoft is also now making security a higher priority. In response to customer concerns about the vulnerability of its products, the company has decided to send 9,000 of its elite programmers on courses in writing secure software.

Language barriers

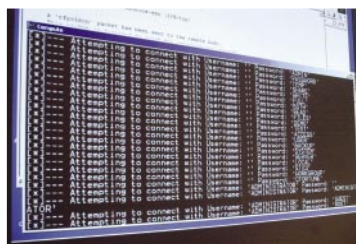
While software vendors try to get their houses in order, what is a sharper focus on computer security likely to mean for the scientist at the bench? In the first instance, unfortunately, it will drain time from research. "Even knowledgeable individuals can find the minutiae of security to be daunting," says one IT official at MIT.

If academic computer networks are to be successfully secured, security experts and scientists may have to take time to find a common language through which to discuss the issues at hand—the jargon of the security professional can be impenetrable to the uninitiated. Observes the IT official from MIT: "More than once, in response to a carefully crafted and technically complete security advisory, we have received feedback along the lines of: 'I'm a professor of chemical engineering, and not stupid, but I've no idea what you are talking about or how to fix the problem you suggest I have.'"

The constant development of software products, and of the tools for hacking into them, means that securing academic computer networks will be an ongoing endeavour. But a new commitment to provide funding for research should help to tip the balance in favour of those trying to protect your data. In February, the US House of Representatives approved the Cyber Security Research and Development Act. Now being considered by the Senate, this would commit some



Fighting back: students at a military college learn how to spot a password attack (left).



\$880 million over five years to create a research programme into computer security to be led by the National Science Foundation and the National Institute of Standards and Technology.

New approaches are likely to recognize that security must be more automated. The idea is to anticipate unknown forms of attack by blocking unusual activity on the network, or reporting it to system administrators. Some researchers, for instance, are developing programs that mimic the function of the immune system to react to both known and unknown threats (see *Nature* **415**, 468–470; 2002).

In the long run, improving security will mean redesigning systems with security uppermost in mind, says Nesbitt. Changes in the way research is done may help to drive this process. In future, many scientists will organize themselves in large international collaborations using 'Grid' networks, which aim to provide supercomputing power on tap by distributing tasks and data over tens of thousands of machines at institutions worldwide.

Grid computing takes security beyond the boundaries of individual institutions, as each must be able to trust the security provisions of others. For example, researchers will need to be able to 'sign in' and then, depending on their level of authorization, access different levels of resources at multiple institutions around the globe, without further authentication. "Succeeding will require rethinking infrastructure, procedures and trust relationships," says Ian Foster, a leader in Grid development at the Argonne National Laboratory near Chicago.

Whether or not you need to use Grid computing, the threat posed by malicious hackers is real, and growing. If you're still in the habit of deleting security advisories from your IT department as soon as they arrive in your e-mail, you may be playing Russian roulette with your data. And if your IT people haven't yet raised security issues with you, perhaps it's time to give them a call. ■

Declan Butler is *Nature's* European correspondent.



Mark Gerstein: lab time is being wasted.

Classic detective puts his finger on the clue

A fictional gumshoe was first to pick up the intriguing criminal potential of 'gum my fingers'.

Sir — Your News report on the fake fingerprints made by mathematician Tsutomu Matsumoto, using gelatin and a plastic mould (*Nature* **417**, 676; 2002), is interesting, but hardly novel. An analogous but more elegant technique was described by the great British mystery writer R. Austin Freeman in *The Red Thumb Mark*, first published by Collingwood Brothers, London, in 1907.

In the book, Freeman, in the guise of his scientific detective Dr Thorndyke, explained how regions of a plate of "chromicized gelatine" exposed to light under a negative image of a fingerprint

would be inversely soluble in hot water, according to how much light each area received. Gelatin under the fingerprint ridges would receive more light and would be preserved when the plate was rinsed with hot water. A three-dimensional gelatin replica of a fingerprint would result.

"The process that I have described," said Thorndyke, "is, in all essentials, that which is used in the reproduction of pen-and-ink drawings, and any of the hundreds of workmen who are employed in that industry could make a relief block of a finger-print with which an undetectable forgery could be executed." I have no

doubt that Freeman's gelatin fingerprint could fool an optical scanner as well as Matsumoto's.

Many more details are provided in *The Red Thumb Mark*, which mystery buffs among *Nature's* readers will doubtless enjoy — even though I have given away part of the ending. This is one of several Dr Thorndyke stories that have recently been reprinted (Stratus, London, 2001).

David Ehrenfeld

Department of Ecology, Evolution, and Natural Resources, Cook College, Rutgers, the State University of New Jersey, 14 College Farm Road, New Brunswick, New Jersey 08901-8551, USA

Microcredit tackles both poverty and birthrate

Sir — It is worrying that population growth should be absent from the agenda of the forthcoming United Nations World Summit on Sustainable Development, as outlined by Wolfgang Lutz *et al.* in Correspondence (*Nature* **418**, 17; 2002).

It is also worrying that empowerment of women can be mentioned without pointing to successful enterprises such as the Grameen Bank of Bangladesh, which makes small loans to help penniless women set up in business. These are known as 'microcredit' or 'microlending' enterprises. To read some of the numerous success stories arising from these small loans, often \$100 or less, see, for example, www.greenstar.org/microcredit or read *Banker to the Poor*, the autobiography of Grameen Bank founder Muhammad Yunus (with Alan Jolis; Aurum, London, 1998).

Microlending, under the rules pioneered by the Grameen Bank, has turned the empowerment of women and the reduction of fertility from mere slogans into a growing reality.

Michael E. McIntyre

Centre for Atmospheric Science, Department of Applied Mathematics and Theoretical Physics, Cambridge University, Silver Street, Cambridge CB3 9EW, UK

Israeli terms are used for consistency, not politics

Sir — Knut Rognes in Correspondence (*Nature* **417**, 379; 2002) accused the *Israel Journal of Entomology (IJE)* of being political, not scientific, claiming that the editor (I. Y.) and a member of the editorial

board (A.F.) refused to accept the geographical terms he used for Israel and adjacent areas in a manuscript he intended to submit (see also *Nature* **418**, 273; 2002).

The policy of the *IJE* has always been to promote science, not politics. One aspect of good science is a precise and consistent terminology. It is particularly difficult to maintain consistent geographical terminology in an area such as that covered by our journal, because of the prevalent and concurrent use of several languages (such as Arabic, Hebrew and English) for the names of the same localities, and the existence of many spelling variants. To overcome this difficulty, the *IJE* established a standard specified in the Notes for Authors (latest version July 2001), as follows: "Names of localities in Israel will be given as they are transliterated in the Israel Touring Map" and "Regions in Israel and nearby areas should follow the *Fauna Palaestina* map". These two maps are given for convenience of scientific purposes only, without any political connotation.

Not only has Rognes tried to impose on us his irrelevant political views, but these views are literally wrong. At present, there is no Palestinian state. If and when a Palestinian state is established, with defined boundaries and scholarly official maps, we will use their transliterated spelling for the appropriate localities. Until then, we will remain with our current policy, as would any scientific journal (including the journal *Fauna of Saudi Arabia*, which, as correctly mentioned by Rognes, ignores the existence of recognized states).

The *IJE* has never "banned any mention of Palestinian national territory", as Rognes claims; as a scientific journal we simply do not find it necessary to deal with political concerns of authors. We hope

that, following this clarification, Rognes will return to us, and we shall welcome his paper in the *IJE*, where it indeed belongs.

Amnon Freidberg*, Ilan Yarom†

*Department of Zoology, Tel Aviv University, Tel Aviv 69978, Israel

†Hazeva R&D Center, Ben-Gurion University of the Negev, PO Box 653, Beer Sheva 84105, Israel

Bioethics: centres reveal sponsors but not policy

Sir — I would like to correct a statement attributed to me in the News story "Code of conduct for bioethics branded 'soft' on corporate ties" (*Nature* **417**, 885–886; 2002). Your article states that the University of Pennsylvania's centre is the only one to provide information about its funding sources on its website.

On 7 June, the Center for Science in the Public Interest wrote to more than 125 bioethics centres and journals, urging them to develop or strengthen their conflict of interest policies. Our survey indicates that although only one centre — the University of Pennsylvania's Center for Bioethics — provides its policy on external funding and conflicts of interest on its website, several others do publicly disclose information about their funding sources.

Although we believe that bioethics centres, individual bioethicists and journals are doing an extremely poor job of disclosing potentially biasing financial ties, they are not quite as negligent as your article suggests.

Virginia Ashby Sharpe

Center for Science in the Public Interest, 1875 Connecticut Avenue, NW, Suite 300, Washington, DC 20009-5728, USA
www.integrityinscience.org

DNA testing for all

There are two fair possibilities for forensic DNA testing: everyone or no one.

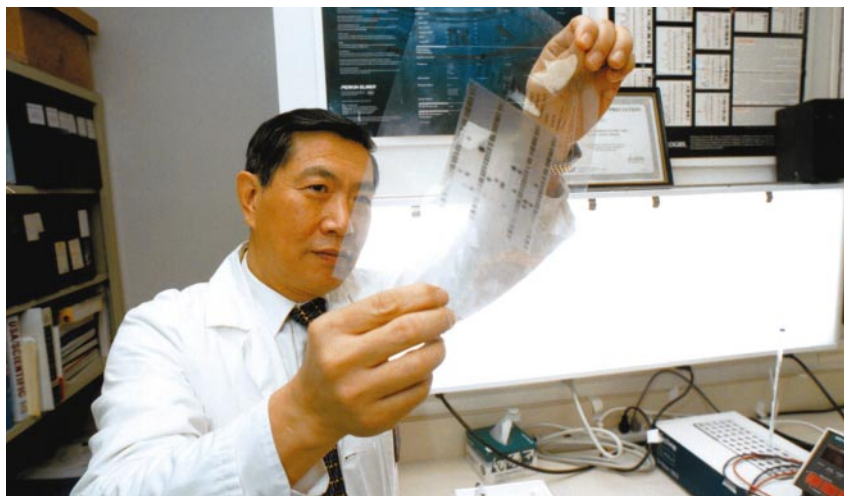
Robert Williamson and
Rony Duncan

DNA testing is the most important advance in forensics in our generation, and probably in the whole of human history. Whether convicting the guilty or acquitting the innocent, a DNA sample is the ultimate form of evidence. But if DNA testing is so valuable, why are we not testing everyone at birth, or at school, or when registering to vote or acquiring a driving licence? The answer is simple: we are scared. Genetics involves the things we really care about — ethnicity and race, sex and family, parents and children. The fears associated with DNA testing run so deep, they simply cannot be ignored.

The police want to catch criminals, and the public wants criminals to be caught. Nobody wants to be wrongly convicted, whether by mistake or through being set up for a crime that they did not commit. The public also wants to know that information about individuals' genes will be kept private^{1,2}. We believe safeguards can be put in place that will protect civil liberties and allay public concern, while allowing police forces use of this extremely powerful technology. The key safeguards are that DNA samples are not retained and that the DNA data are placed on a database that is independent of the police. More important, we believe that there are only two possible fair practices when it comes to DNA testing for forensic purposes. Either test everyone, or test no one. We argue that the former is the smarter option.

The facts

DNA testing relies on the fact that each human has a unique DNA sequence, and that this sequence can be used to link a person to — or eliminate them from — any crime



A database of DNA samples could be invaluable for forensics, but it would need careful regulation.

scene where a sample of human tissue for DNA analysis has been left. DNA tests have become extremely sensitive during the past 10 years. Gene amplification techniques now allow a unique DNA fingerprint to be obtained from a single nucleated blood cell, cheek epithelial cells from saliva, sperm, a single hair follicle or the skin cells from a fingerprint (see Box 1).

It is difficult to give those who do not work with DNA a realistic feel for the exceptional power of these tests. It has been estimated that if 10 DNA sites (or loci) are tested, the chance of a random match between two people is one in a billion³.

DNA forensic databases have been set up by police forces around the world. The protocol for DNA testing differs from country to country and from state to state. The United Kingdom has the most liberal stance — samples are taken from anyone who is suspected

of, charged with, reported for or convicted of a recordable offence⁴. The United Kingdom is aiming to hold the DNA profile of nearly 1 in every 15 people in Britain⁵.

The fears

It was once suggested that members of the police in Tasmania should voluntarily submit DNA samples from themselves to allow the elimination of any police tissue that might contaminate crime scenes. Despite — or perhaps because of — their close knowledge of DNA testing, they refused to undertake testing on the grounds that it would infringe their civil liberties⁵. The fears about this practice are clearly not a consequence of ignorance, but extend throughout the community, including those most knowledgeable about the technology's power. So what are we most scared about?

There are two fundamental aspects to the fears surrounding DNA testing. First, there are general fears that DNA data will be used to violate privacy and that the government, insurance companies, employers, colleagues or family will access genetic information that we, as individuals, don't even have ourselves (let alone understand). People are worried that the database will be used for paternity tests or genetic research. The second fear is related to wrongful convictions. People are frightened of being wrongly convicted as a result of contamination or even an error in the testing process, and they are petrified of being set up for a crime they did not commit.

It is crucial that specific safeguards are introduced to protect the community and address these fears. These safeguards must address all four aspects of the testing process:

Box 1 The PCR technique

The polymerase chain reaction (PCR) is an experimental technique that amplifies a specific sequence of DNA, usually from a few hundred to a few thousand base pairs in length, using a sample of an individual's cells.

Each of us, apart from identical twins, differs at several million places in our DNA sequence from any other human being. Because 'scene of crime' samples often contain blood, sperm or other body tissue, the personal genetic fingerprint, or profile, is easily obtained when the amplified DNA bands are separated and sequenced.

If the same short DNA sequences are amplified

by labs in all countries, it will be easy to create an international 'identity' database. Most variation occurs in non-coding sequences, so it will be simple to avoid data giving information about the appearance or phenotype of any individual.

If run by experienced technologists, PCR is free of artefacts, and contamination cannot occur. What can happen is identification of a person who was present at the crime scene, but who was not involved. It is therefore important to be explicit about the type of DNA sample in a court of law. A jury may, for example, give more weight to a sperm sample than to a saliva sample from a cigarette butt.

(1) the preservation and handling of DNA samples taken from crime scenes; (2) access to DNA samples taken from convicted offenders, suspects or the general community; (3) accreditation of the laboratories carrying out the analysis (and the scientists working in these laboratories), and access to independent expert advice by the defence, to check that no contamination or error has taken place; and (4) access to and storage of the DNA profiles that are the result of the analysis (see Box 2). It is also critical that any database used for criminal-justice purposes must not be used for any other purpose (such as research), or the public will lose confidence in its independence. Once people's fears are allayed, the real issue, and perhaps the most controversial aspect of DNA testing, is the question of whom to test.

All or nothing

Let us assume that credible safeguards are created and maintained, and that DNA databases are set up across the world with checks in place. From whom should DNA samples be taken? There are three possible alternatives. First, samples could be taken only from convicted criminals — and perhaps even for specific crimes, such as crimes of violence. Second, samples could be taken not only from convicted criminals, but also from suspects and anyone who comes under suspicion and is willing to give consent. Third, samples could be taken from every person in the community, at birth, for example.

Taking samples only from convicted criminals would send a message to the public that, as members of our society, individuals have the right to keep their genetic profile private. But if convicted of a crime, a person loses that right. In other words, the taking of a DNA profile acts not only as a punishment, but also as a deterrent — at least in theory.

The concept of DNA testing as a punishment has negative connotations and is likely to increase public concern. If deterrence



Sample case: a DNA test showed that Thomas Webb (centre) had been wrongly convicted of rape.

from future crime is the true motivation in taking DNA samples from convicted criminals, why would we not want to deter all members of society from criminal activity by collecting DNA samples at birth? The possibility of sending a message about social and racial inequality is another concern with testing only those who have been convicted of a crime because, in many societies, members of minority ethnic groups are convicted at a higher rate than other members of the community and are overrepresented in the prison population⁶.

If samples are to be taken from the convicted, from suspects and from anyone who volunteers to have a sample taken, the rules of consent become clouded. If individuals

refuse to have a sample taken, the immediate assumption is that they have something to hide. It has therefore been suggested that, because a failure to consent could raise suspicion, such consent is not truly voluntary⁷.

The most logical and fair practice — and also the most controversial — would be to DNA-test all individuals at birth. This would not only act as a deterrent from crime for all members of the community, but would make the task of catching criminals easier for police. If the correct safeguards are in place to protect civil liberties, why should a proposal to test everyone at birth be a frightening one? On the other hand, if the correct safeguards are not in place, and the fears are justified, why are we daring to test anyone at all? ■

Robert Williamson and Rony Duncan are at the Murdoch Children's Research Institute, University of Melbourne, Royal Children's Hospital, Flemington Road, Parkville, 3052, Victoria, Australia. Correspondence should be addressed to Robert Williamson, e-mail: williamsb@cryptic.rch.unimelb.edu.au

Box 2 Safeguards for forensic DNA databases

We propose the following five measures to safeguard civil liberties and allay public concern. Of primary importance is the independence of both the laboratories that test DNA samples and the DNA database itself.

1. Forensic laboratories responsible for DNA analysis should be independent of the police, and should be fully accredited using a national scheme, including definition of acceptable staff qualifications, periodic assessment and an enforceable code of conduct.
2. Any DNA sample taken from a crime scene should be sent directly to the laboratory for storage and testing. Where possible, it should be divided, and one portion should be reserved

for the defence. Samples should always be kept in tamper-evident packaging.

3. Blood, cheek swabs or other samples from suspects, convicted offenders or volunteers should be sent directly to the laboratory for testing, and not be placed in police custody. These samples should be destroyed after analysis has taken place and DNA profiles entered as data.
4. All DNA profiles should be computerized and held on a national or international DNA database that is independent of police. The database needs to be accessible by police for effective use, but it is critical that police cannot enter or alter any data.
5. Any data accessed from a DNA database outside this accredited, independent process should not be admissible as evidence.

1. Nowlan, W. *Science* **297**, 195–196 (2002).
2. Rothenberg, K. H. & Terry, S. F. *Science* **297**, 196–197 (2002).
3. Concar, D. *New Scientist* **10** (5 May 2001).
4. Australian Law Reform Commission and Australian Health Ethics Committee *Protection of Human Genetic Information: Issues* (2001).
5. Police Association of Tasmania Newsletter (March 2002).
6. Reilly, P. *Nature Rev. Genet.* **2**, 313–317 (2001).
7. Kimmelman, J. *Nature Biotechnol.* **18**, 695–696 (2000).

Acknowledgements

We thank Julian Savulescu (Murdoch Children's Research Institute) and James Robertson (Australian Federal Police) for helpful comments, but the views expressed are those of the authors only.

Putting Egypt back on the map

Ahmed Zewail's autobiography looks at the future of science in his homeland.

Voyage Through Time: Walks in Life to the Nobel Prize

by Ahmed Zewail

American University in Cairo Press: 2002.

256 pp. \$22.95

A. B. Zahlan

When Ahmed Zewail won a Nobel Prize in 1999 he was the first Arab scientist to do so, and only the second Muslim, behind the Pakistani physicist Abdus Salam in 1979. Zewail's greatest contribution as a chemist, for which he won the Nobel Prize, was his pioneering study of rapid chemical and physical processes through the use of short laser flashes. A flash lasting 20–60 femtoseconds (one femtosecond is 10^{-15} s), which is used to excite a molecule, is followed by a second pulse, at a different wavelength, to monitor the state of the system at well-defined intervals. Zewail developed this method for studying chemical transformations as they occur. The duration of a molecular vibration is around 300 femtoseconds.

Voyage Through Time is Zewail's autobiography, and covers his education and personal life, and his aspirations for the future of Egypt and more particularly for science in his native country.

Zewail was born in 1946 in Damanhour, a town with a population of 200,000 in Egypt's Nile Delta. After primary and secondary education in the village school of neighbouring Desuq, he attended the University of Alexandria, where he obtained a BSc in chemistry in 1967, followed two years later by an MSc in the same field. He excelled at the university, and before long travelled to the United States, where he had been awarded a fellowship by the University of Pennsylvania. In 1973 he earned his PhD at the University of Pennsylvania. *Voyage Through Time* provides a cogent account of Zewail's adjustment to his new environment, in which he was to flourish. His productive and creative abilities were soon recognized and rewarded in his adopted country.

After graduation, Zewail obtained a postdoctoral position at the University of California, Berkeley. From there, he moved in 1976 to Caltech, where he held his first permanent academic post. In 1990 he was appointed to the prestigious Linus Pauling chair of chemical physics at Caltech. Throughout his distinguished academic career he received a series of awards that finally led to Stockholm and the Nobel Prize in chemistry.

What were the factors that shaped the formative years in Egypt of this gifted young scientist? This question remains largely



Bright prospect: Ahmed Zewail wants to see the creation of a world-class science university in Egypt.

unanswered in *Voyage Through Time*. We are told surprisingly little about his formative experiences as a student in a country that was undergoing radical transformations throughout his youth. By contrast, in later chapters, Zewail talks expansively about his hopes for the development of science in Egypt. But we are not told, for example, how his experience at the University of Alexandria contributed to his achievements. Instead there are only loose references to historical figures such as Alexander the Great and Lawrence Durrell, neither of whom could have made a significant impact on the author's intellectual formation.

The 1950s and 1960s, when Zewail was growing up, were decades of momentous change in Egypt. Although there is passing reference to the role played by President Gamal Abdel Nasser in providing widespread access to higher education, which in turn enabled Zewail to attend the University of Alexandria, there is little else to convey the impact of those years on the author. There are no insights into either the intellectual life of Egypt at the time or the monumental changes — except for the democratization of the educational system — that were taking place.

The book tells us that Zewail has now visited all of the major research institutions in

Egypt, yet we are not given any comments on their status, their strengths and weaknesses. Such a book would have provided an ideal opportunity to initiate a serious and frank public discussion on the state of science in Egypt. Furthermore, that kind of discussion would be a vital step in the preparation for plans that Zewail proposes for Egypt.

Zewail seems to have set his heart on the creation of an Arab Science and Technology Foundation (see *Nature* **416**, 120–122; 2002) with the mission of establishing a University of Science and Technology and an associated Technology Park (see *Nature* **410**, 741; 2001). The project has the support of President Hosni Mubarak, and 300 acres of land have already been designated to serve as the university's campus.

But I doubt whether a world-class university can be established in Egypt without changes in the political culture. In the past 50 years, some 200 new universities have been created in the Arab world, none of which is even on the way to becoming world-class. More than 100,000 Arabs (including a high proportion of Egyptians) have earned advanced degrees at prestigious institutions in the West during this period; some 60% of these have returned to their home countries but remain strikingly marginalized. Moreover, Egypt has witnessed a series of attempts

since 1820 to develop its technology base; it is difficult to point to any that have achieved the hoped-for results. These and other facts need to be addressed before any realistic plans for change are made. ■

A. B. Zahlan is a science policy consultant, formerly professor of physics at the American University of Beirut, and director of the Royal Scientific Society, Jordan.

Quantifying bird diversity

Evolutionary Ecology of Birds: Life Histories, Mating Systems and Extinction

by Peter M. Bennett & Ian P. F. Owens
Oxford University Press: 2002. 292 pp.
£49.50, \$85 (hbk); £24.95, \$45 (pbk)

Arie J. van Noordwijk

Birds vary enormously. In David Lack's 1968 book *Ecological Adaptations for Breeding in Birds* (Methuen), he explained why some birds are very long-lived and others have high mortality, and why some birds lay two dozen eggs each year whereas others lay just one egg every two years. The new book by Peter Bennett and Ian Owens, *Evolutionary Ecology of Birds*, claims to be the most comprehensive reappraisal of avian diversity since then.

Whereas Lack had to rely on reasoning and careful selection of examples, Bennett and Owen have used the comparative method to analyse variation among species, families and orders. The big problem in analysing observational data on birds is that, for example, the fact that many passerine birds are small, have high mortality and high fecundity may be due either to selective forces or to common history. So, in a statistical sense, the data on all songbirds, which make up more than half of all bird species, cannot be regarded as independent data points.

The comparative method is based on locating contrasts in a phylogenetic tree — a map of relationships between species. Each branch (or node) within the tree represents a single evolutionary event leading to the generation of two new groups (or species) from a single ancestral group. We can regard this change as one unit and investigate which traits differ by how much at the same node in the tree. We can then sum all these changes over all instances where we have found the same evolutionary transition. This process removes common history as a source of covariation. Providing that the taxonomy is a correct representation of evolutionary history, this method allows a rigorous testing of hypotheses.

This book illuminates the application of



Enough is enough: like many birds, American robins lay clutches of just a few eggs at a time.

the comparative method to questions about the life histories, mating systems and extinction risks of birds. However, the comparative method has been applied only to around one-third of the questions described in the book, which makes for a rather unbalanced, although thought-provoking, account.

Has progress been made since 1968? It certainly has in the area of mating systems, and in the analysis of extinction risk. There has also been progress in the understanding of life-history variation, but I am not entirely convinced by some of the results reported here. For example, the authors identify a major division between 'fast' (high reproduction, low adult survival) and 'slow' lifestyles, which is related to having protected nests or not. But, in my view, the lumping of safe breeding sites (hole nesting or colonial breeding) leads to unconvincing results. Predation can be quite high in nesting holes but, more importantly, there is not a single mention of parasites as a major selection force on nesting habits.

By contrast, the splitting of patterns in coloration according to the type of pigment involved, and the splitting of extinction risks into those resulting from habitat loss and those caused by increased mortality rates, are convincing. In these sections, the power of deducing processes from patterns is nicely illustrated.

The section on life histories begins with a report that nearly all variation is at the family or order level, rather than at the species level. I work on species with more than twofold variation in reproductive and mortality rates, so this seems a puzzling observation. Both here and in the section on mating systems, the authors derive a hierarchical explanation of the observed variation. Different processes are claimed to be responsible for the observed patterns at different taxonomic levels. I find this unattractive. What I

miss in these discussions are considerations of the potential rate of change in the traits involved. I am convinced that twofold changes in reproductive rate can be achieved in the order of 10–100 generations in nearly all species.

In my view, the absence of considerable variation at lower taxonomic levels should be attributed to the absence of net selection forces or, in other words, to ecological



Flocking together: large numbers of carmine bee-eaters congregate to nest in cavities in cliffs.

J. FOOT/NATUREPL.COM

K. SCHAFER/CORBIS

equilibria. The question of whether the absence of variation at lower taxonomic levels is due to the absence of net selection or the absence of genetic variation to respond to selection is crucial in evaluating extinction risks in a world where important elements in the environment of all species are changing at unprecedented rates. Over the past two centuries, many bird species have radically changed many aspects of their habitat use, life histories and mating systems. The absence of variation at the species level in the analysis of major life-history components tells me either that the database is wrong or that far fewer changes have occurred than would have been possible. I think that a more plausible explanation for the absence of net selection for major life-history changes is that adjacent niches were already filled by other taxonomic groups.

In summary, this book is a progress report. Almost 35 years have passed since Lack's epic work was published. Nearly all of the progress in our understanding of avian diversity has been made in the past decade. This book provides a good record of this progress but I hope and expect that it will be outdated a decade from now. Meanwhile, it provides a useful overview of the strengths and weaknesses of the comparative method, and it certainly stimulates thinking about how we can learn more about variation in avian life histories and mating systems. Part of this stimulation is achieved through annoying the reader. Whether or not this was intended doesn't really matter — it works. ■

Arie J. van Noordwijk is at the Netherlands Institute of Ecology, PO Box 40, 6666 ZG Heteren, The Netherlands.

Thinking laterally about genes

Lateral DNA Transfer: Mechanisms and Consequences

by Frederic Bushman

Cold Spring Harbor Laboratory Press: 2002. 448 pp. \$59, £43 (hbk); \$39, £28 (pbk)

W. Ford Doolittle

Lateral (also known as horizontal) gene transfer (LGT) is gene exchange across species boundaries, and is a hot topic in some circles. Microbial genomicists and phylogeneticists in particular become more excited about it with each new prokaryotic genome sequence that reaches completion. No one had anticipated that prokaryotic genomes would be so extensively mosaic, so obviously pieced together by the assembly of genes from so many different antecedents, in some never-ending game of molecular Lego.

The genomicists are generally happy about LGT. The acquisition of foreign genes



A monumental study

Studying monuments such as this zoomorph found at Quiriguá in Guatemala became much easier after Alfred Maudslay's visits to the ruined cities of Mesoamerica. Maudslay took many photographs but also arranged for plaster casts to be made, enabling scholars to study these artefacts of Maya culture in relative comfort away from their jungle home. Maudslay's careful documentation proved invaluable in the deciphering of Maya hieroglyphics. Ian Graham's book *Alfred Maudslay and the Maya* (University of Oklahoma Press, \$39.95) is the first biography of this pioneering archaeologist, and contains many of Maudslay's own photographs.

and gene clusters can explain some of the striking differences in the properties of bacterial species (or strains within species) that by other criteria are almost identical. Large (>20%) differences in genetic make-up between pathogenic and benign strains of the same species (*Escherichia coli* O157:H7 and K12, for instance) can best be understood with models that invoke evolutionarily rapid gene acquisition (by LGT) and loss. But phylogeneticists are less pleased. They want to use gene sequences to reconstruct evolutionary trees for species on the assumption of 'vertical' gene transmission between generations, but if genes can switch species, these trees become webs — much messier. So when LGT has been inferred from comparative data only, they often somewhat petulantly ask: "What's the mechanism?"

Frederic Bushman's book, *Lateral DNA Transfer*, doesn't attempt to answer this challenge for all inferred instances of LGT. But by providing reasonably thorough and unusually readable accounts of current knowledge about the many ways in which DNA can be exchanged within and between both prokaryotes and eukaryotes, Bushman has provided LGT's proponents with a broad range of reasonable rejoinders. And he has given us even more confidence that, where there is a will (a selective advantage) for gene exchange, there is a way.

In 12 chapters Bushman summarizes the states of play in work on conjugation, transposition, transduction, antibiotic resistance (mechanisms and spread), pathogenicity

islands, comparative prokaryotic genomics, retrovirology (with a special chapter on HIV), retrotransposons (LINEs and SINEs), Tc1/Mariner, P and other DNA elements, eukaryotic genomic composition and fluidity, the vertebrate immune system, and inter-domain transfer (Ti plasmids in particular). These topics are not usually treated together, and are usually presented to students only after they learn about 'real' genes. So Bushman is opening the door to a more integrated general molecular evolutionary discipline that would treat all mobile elements (and in the extreme, all transferred genes) as evolutionary agents in their own right, surviving not only because of benefits that they confer on the genomes in which they currently occur, but because they are especially transferable between genomes.

The 'selfish operon' hypothesis of Jeffrey Lawrence and John Roth is the best articulated and most radical theory that this new discipline currently has. Bushman endorses it, but I wish that he had presented this theory (and several other evolutionary mechanisms) in rather more detail.

Bushman does attempt a broader integration in his final two chapters, the first of which deals with regulation (usually negative) exerted on LGT by the recipients and agents of transfer. Mostly this has to do with limiting the copy number of mobile elements within genomes or the virulence of viruses, rather than with modulating interspecies exchange, and what Bushman has to say about the latter seems less sophis-

ticated. He cites but then slights Ichizo Kobayashi's elegant selfish-element theory to explain restriction modification systems, and misses out completely on Rosemary Redfield's data on transformation that show that bacteria care most about the food value of foreign DNA.

Similarly, the final chapter summarizes Bushman's earlier generalities about the differences between prokaryotes and eukaryotes with regard to the mechanism, frequency and impact of LGT, and other people's ideas

about the role of LGT in the evolution of introns, sex and life. The bold new unifying theory I had hoped to see simply isn't here, and in particular there is no attempt to bridge the gap to population genetics: this is a molecular biologist's book. But then I could do no better myself, and Bushman's book does provide, in one place, the best review of the many pathways and processes of LGT now available. I'm glad it's on my shelf. ■

W. Ford Doolittle is in the Department of Biochemistry and Molecular Biology,

Dalhousie University, Halifax, Nova Scotia B3H 4H7, Canada.

More on gene transfer

Mobile DNA II

edited by Nancy L. Craig, Robert Craigie, Martin Gellert & Alan M. Lambowitz
ASM Press, \$169.95

Horizontal Gene Transfer, 2nd Edition

edited by Michael Syvanen & Clarence I. Kado
Academic Press, \$99.95

Science in culture

Algorithmic architecture

The Serpentine Gallery Pavilion by Toyo Ito and Cecil Balmond.

Martin Kemp

The most spatially ambitious buildings have invariably extended the potential of engineered structures into new realms. We need only think of the soaring Gothic cathedrals, the ultimate high-tech of their day. Mathematical principles are integral to this quest, even when the methods have been more empirical than theoretical.

At key moments — as with Antoni Gaudí, architect of the church of the Sagrada Família in Barcelona, and Richard Buckminster Fuller, who invented the geodesic dome — the mathematics of structure has been consciously embedded in the broader context of the generation of form within nature. Rather than the literal imitation of natural structures, this ambition exploits design processes that are analogous to nature's organizational principles.

For like-minded contemporary architects, designers and engineers, the advent of the computer has released radical new potentiality, as witnessed in the joyous pavilion erected outside the Serpentine Gallery in London's Hyde Park by the Japanese architect Toyo Ito and the London-based engineer Cecil Balmond of Ove Arup. To describe them respectively as architect and engineer is inadequate. Both are visionaries who blend structural insight, innovative processes of form generation, aesthetic adventure and wide cultural reference. They are forging new modes of building that subvert the dominant box of modernist architecture.



Ito's personal vision is to transmute traditional buildings, viewed as 'bodies', into entities adapted to the electronic age. He argues that the "virtual body of electron flow is drastically changing the mode of communication in family and community, while the primitive body in which water and air flow still craves for beautiful light and wind. The biggest challenge for us is how we can integrate these two types of body."

Balmond, for his part, is fascinated by proportion and number theory, and has published a tale of numerology, *Number 9: The Search for the Sigma Code* (Prestel, 1998). He delves into pythagorean harmonics, sacred geometry, Islamic tiling, tantric numbers, the mathematics of symmetry and asymmetry, chaos theory and fractals for interlocking insights into the magic of form and number.

It is perhaps surprising to find that the underlying form of the pavilion is so simple. It is a cubic structure, 17 metres square and 4.5 metres high. But this is where the simplicity ends. The floor and walls dissolve into an intricate web of interpenetrating squares, triangles and irregular polygons generated by an algorithm. From the mid-point of one side of a square, a line is drawn to a point two-thirds along an adjacent side. The procedure is then repeated to produce an internally 'suspended' square which is rotated

on the diagonal axis. Through iteration and extension of this linear construct, a flat sheet is generated, which folds perpendicularly to produce the walls.

The not inconsiderable load of the roof glides above our heads, like a flight of tethered kites. The angular music of the design, with its alternating sets of opaque and open panels, generates a visual dynamism that subverts architectural convention. The net impetus is up and across rather than static, as in the classic post-and-lintel system.

Building a temporary structure of this kind, as part of the Serpentine Gallery's annual programme, has proved liberating for Ito: "One need not be so strict about function nor worry about how it will age. And, it seems to me, it just might offer the clearest expression of the concepts I habitually imagine." The pavilion serves, in effect, as a laboratory for a structural aesthetic which would have been inconceivable in an earlier era.

Martin Kemp is in the Department of the History of Art, University of Oxford, Oxford OX1 2BE, UK.

Visualizations: *The Nature Book of Art and Science is a collection of essays edited by Martin Kemp (published by Oxford University Press and the University of California Press; £20, £35).*



Diminishing returns

Henrik G. Dohlman

On sunny days I often slip out of the lab and visit a nearby cafe, where I am greeted by the aroma of freshly brewed coffee and oven-baked pastries. Within a few minutes, however, these odours become less noticeable. They're not gone, but my ability to perceive them has somehow diminished. My typical order, a large coffee with a shot of espresso, no longer provides the nervous jolt it once did — an unfortunate consequence of my caffeine habit. Staring into the bright green lights above makes the white pages of *Nature* appear slightly orange. These familiar experiences are all manifestations of desensitization — the underlying mechanisms of which have more in common than might be expected.

Caffeine, odours and light, as well as most hormones and neurotransmitters, act on a family of cell-surface receptors that span the cell's membrane seven times. Several of these 'serpentine' receptors bind to useful drugs (such as β -blockers and antihistamines) and also to some illicit ones (such as LSD and cannabis). Even in the simplest eukaryotic organisms, related receptors respond to mating pheromones, chemoattractants and nutritional signals. In each case, receptor stimulation leads to the activation of a G protein, which binds to the nucleotide GTP and transmits a signal to an effector enzyme inside the cell. Signalling persists until GTP is converted to GDP, when it stops.

In general, desensitization is characterized by feedback inhibition after a delay. Any external stimulus that triggers a response inside the cell also triggers a process designed to inhibit the response to further exposure to the same stimulus. Desensitization can occur on the order of seconds

(to a flash of light), minutes (to odours) or even days (to caffeine).

Receptors have long been considered as likely targets of desensitization, a logical hypothesis considering their role as the 'gatekeepers' of the cell. Prolonged stimulation of most receptors leads to rapid phosphorylation, physical uncoupling from the G protein and internalization. The initial phosphorylation event can result from conformational changes in the agonist-bound receptor, because this also makes it a better substrate for some kinases. For instance, activation and phosphorylation of the green photoreceptors can temporarily alter colour perception, making the world appear less green and more orange. Alternatively, phosphorylation can result from activation of a downstream effector kinase that has among its substrates the activating receptor. Long-term desensitization can occur through increased expression of the relevant kinase, or reduced expression of the receptor. There is ample precedent for each of these suggestions.

Desensitization can also occur at later stages of the pathway — G proteins and effectors are also targets of desensitization. About five years ago, a family of proteins, known as RGS (for 'regulators of G-protein signalling') proteins, was discovered in yeast and was subsequently shown to promote desensitization by accelerating GTP hydrolysis by G proteins, thereby rapidly inactivating the signal. In at least some cases, RGS expression is strongly induced by activation of the pathway, as is typical of feedback-inhibition loops.

More recently, attention has turned to other types of protein modification besides phosphorylation. Some common biochemical modifications, such as glycosylation and myristoylation, are fairly static and are therefore unlikely to have a role in desensitization. The activation of some receptors, RGS proteins and at least one effector kinase result in ubiquitination of these proteins — the covalent attachment of multiple ubiquitin polypeptides to the substrate protein. These proteins are usually degraded quickly by a complex of proteases called the 26S proteasome. In yeast, ubiquitination of the receptor requires prior phosphorylation. This situation is unusual, however, in that the yeast receptor is mono-ubiquitinated rather than poly-ubiquitinated, and is delivered to the vacuole compartment rather than to the proteasome.

Another surprise is the discovery that the downstream kinase Ste7 is also ubiquitinated in response to a chronic stimulus. Ste7, a member of the mitogen-activated protein

Desensitization

Any external stimulus triggering a response inside the cell also triggers a process inhibiting the response to further exposure to the same stimulus.

kinase (MAPK) kinase family, is thought to be a limiting component of the MAPK cascade of reactions. There is now good evidence that the ubiquitination of Ste7 is part of a feedback-inhibition loop that leads to MAPK desensitization. So, whereas attention in the past has focused on desensitization through receptor phosphorylation, we can now consider a new paradigm of desensitization of G proteins and effector enzymes, through modifications other than phosphorylation.

History shows that breakthroughs in our understanding of cell regulation are often followed by an increased understanding of counter-regulatory events. Consider the tandem discoveries of protein kinases and phosphatases, or oncogenes and tumour-suppressor genes. The discovery of RGS proteins, acting in opposition to cell-surface receptors, is another example. What will be next? I believe that the current focus on cell regulation by ubiquitination will soon shift to mechanisms of protein de-ubiquitination. In yeast, there are 16 ubiquitin-processing proteases. This large number suggests that there are highly specific functions for each family member. Deletion of the gene for just one such protease, *UBP3*, enhances the pheromone response in yeast and promotes pheromone-dependent ubiquitination of the effector kinase Ste7. Perhaps other family members will have similar, pathway-specific effects on cell regulation.

The existence of various neurotransmitters, receptors, G proteins and effector enzymes has long been part of college textbooks and popular reviews. Feedback inhibition is at last gaining a similar degree of recognition, which (as befits the subject) has occurred after a period of prolonged activity and a suitable delay. ■

Henrik G. Dohlman is in the Department of Biochemistry and Biophysics, University of North Carolina, Campus Box 7260, Chapel Hill, North Carolina 27599-2852, USA.

FURTHER READING

Wang, Y. & Dohlman, H. G. *J. Biol. Chem.* **277**, 15766–15772 (2002).
Shenoy, S. K., McDonald, P. H., Kohout, T. A. & Lefkowitz, R. J. *Science* **294**, 1307–1313 (2001).
Hicke, L. *Nature Rev. Mol. Cell Biol.* **2**, 195–201 (2001).
Ross, E. M. & Wilkie, T. M. *Annu. Rev. Biochem.* **69**, 795–827 (2000).

KOBAL COLLECTION



Force of habit: prolonged nicotine intake causes smokers to become less sensitive to its effects.

Uncertain sinks in the shrubs

Christine L. Goodale and Eric A. Davidson

Replacement of grassland by shrubland, which is occurring on a large scale in the United States, is thought to lock up considerable amounts of carbon. This 'carbon sink' may be much smaller than previously estimated.

Balancing the global carbon budget just got more difficult. For decades, there has been debate over what happens to the carbon dioxide released from the burning of fossil fuels and clearing of tropical rainforests. About half of it accumulates in the atmosphere^{1,2}, prompting concerns about global warming. The ocean and land take up the rest, acting as carbon sinks, and understanding where and why these sinks occur is essential for managing them in the future. Now, new field measurements cast doubt on the estimated size of one potentially large but poorly quantified carbon sink — that of woody shrubs when they encroach into and replace grassland, a process that has been thought to lock up large amounts of extra carbon. As they describe on page 623 of this issue, Jackson *et al.*³ show that there are smaller increases in carbon storage than anticipated, and in some cases actual reductions.

There is general agreement that land in the Northern Hemisphere acts as an important carbon sink. Yet estimates of its exact size, and the specific locations and

contributing factors, vary greatly, due in part to different measurement approaches. Calculations based on measurements of carbon dioxide and oxygen in the atmosphere^{1,2} nearly always produce larger estimated sinks for North America and Eurasia ($0.6\text{--}2.7 \times 10^{15}$ grams — petagrams — of carbon per year) than do those based on ground measurements of forest growth ($0.6\text{--}0.7 \text{ Pg C yr}^{-1}$).

Last year, however, a carbon budget produced for the continental United States partly reconciled top-down (atmospheric) and bottom-up (land-based) methods by including estimates of both forest and non-forest sinks⁴. Together, the bottom-up estimates ($0.4\text{--}0.7 \text{ Pg C yr}^{-1}$) roughly matched those from the top-down approach ($0.7 \pm 0.5 \text{ Pg C yr}^{-1}$). This reconciliation depended on the existence of a rather large sink (about $0.13 \text{ Pg C yr}^{-1}$) in the rangelands of the western United States^{4,5}. The suppression of fires and overgrazing have favoured the expansion of trees and woody shrubs into grasslands⁶, over an estimated area of as much as 220 million hectares in the United States^{4,5}. This 'woody encroachment' formed 18–34% of the total estimated sink for carbon in the continental United States — it was by far the largest non-forest sink, and the least secure term in the land-based carbon budget⁴. Jackson and colleagues' measurements³ add further uncertainty to these estimates.

The authors studied six pairs of grassland and woody-shrubland sites (Fig. 1) along a rainfall gradient in the southwestern United States. They found that conversion of grassland into shrubland increased the amount of carbon in vegetation, but usually only by a small amount. Stocks of organic carbon in soil also increased slightly in the dry sites. In the wetter sites, however, carbon stocks decreased, particularly in the top metre of soil. On net balance, the drier sites gained a small amount of carbon and two of the three wetter sites lost it. These results are consistent with a literature review showing that soil carbon decreases when pastures are planted with trees, with larger decreases occurring at wetter sites⁷. If the sinks investigated by Jackson *et al.*³ are typical of other sites, these results indicate that the sink in the United States caused by woody encroachment^{4,5} was substantially overestimated (Fig. 2).

Why did soil carbon decrease at the wet-

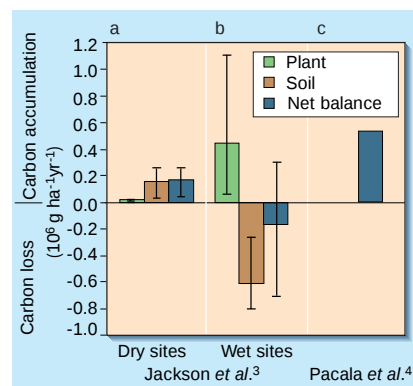


Figure 2 Carbon accumulation and loss due to expansion of shrublands into grasslands.

a, Jackson *et al.*³ report low rates of carbon accumulation in shrubland plants and soils at dry sites in the southwestern United States.
b, The authors³ found that plant carbon increased at wetter shrubland sites in the same region, but so much carbon was lost from the soil that there was a net loss to the atmosphere. Means and ranges are shown for three dry sites and three wet sites. All of these sites had lower rates of carbon accumulation than the presumed mean, depicted in c, for all US non-forest, non-cropland areas reported by Pacala *et al.*⁴. If this higher mean sink due to woody encroachment in the western United States were correct, the US carbon sinks calculated by atmospheric models and by ground-based measurements would appear to be nearly reconciled. The results of Jackson *et al.* cast doubt on that conclusion.

ter sites? In soil, organic-carbon stocks represent the net balance between plant inputs of carbon and subsequent losses through erosion, leaching, and decomposition by microbes. For soil carbon to decrease, either inputs decreased or losses increased (or both). Grasses usually allocate more of their production to roots than to shoots⁸, and most of the roots are in the top metre of soil. Furthermore, grassland productivity correlates strongly with rainfall⁹, so that grassland soils in the wet end of the gradient receive more root input of carbon.

When shrubs invade grasslands, carbon input to the topsoil may decrease because the woody species allocate less of their total production below ground, and what they do send to roots is distributed more deeply⁹. Because much of the carbon in soil decomposes slowly, several decades must pass after



Figure 1 On site — a view of the Sevilleta area, one of the sites where Jackson *et al.*³ conduct their long-term research.

ROB JACKSON

vegetation changes before a new equilibrium is approached, and so Jackson *et al.*³ chose shrublands that were at least 30 years old. The lower soil-carbon stocks in shrublands than they measured at the wet end of the rainfall gradient indicate that inputs from woody invaders during the past 30–100 years were less than losses of old soil carbon from the grasses that previously occupied the site. This process is presumably continuing on lands where woody encroachment is more recent.

On the dry end of the gradient, Jackson *et al.* found little change in soil carbon following woody encroachment, suggesting that there was little change in production below ground. This explanation is speculative, as Jackson *et al.* did not directly measure inputs and losses of soil carbon, a very difficult task. Unfortunately, the authors also did not measure carbon stocks in woody roots (those of 1 cm or more in diameter), and so the sink in deeply rooted woody shrubs is probably an underestimate.

Woodlands, savannas, shrublands and grasslands cover about 40% of the Earth's surface¹⁰, and so their potential role as carbon sinks — or sources — is a key factor in the global carbon budget. Measuring the effects of woody encroachment at particular sites is one challenge; extrapolating the results to regional or larger scales is quite another. Particular sites are certainly large sinks for carbon¹¹, but the global extent of grassland replacement by shrubland is highly uncertain^{3,4}. For a few regions, analyses of historical aerial photographs or satellite data have produced estimates for the extent of encroachment (ref. 12; G. P. Asner, personal communication). Yet the results of Jackson *et al.*³ complicate these assessments by showing that increases in plant biomass above ground may be more than offset by losses of carbon from soil. This cancelling of gains in carbon above ground with losses below ground has already been demonstrated for individual sites¹³, but we now have evidence of a regional phenomenon that appears to vary somewhat predictably along a climatic gradient.

What are needed are more reliable regional estimates. That will require comprehensive assessments of the extent to which shrubs are replacing grassland, along with field measurements and model simulations of the size and variability of plant and soil carbon stocks across a range of climate conditions.

Christine L. Goodale and Eric A. Davidson are at The Woods Hole Research Center, PO Box 296, Woods Hole, Massachusetts 02543, USA.

e-mail: cgoodale@whrc.org

1. Prentice, I. C. *et al.* in *Climate Change 2001: The Scientific Basis* (eds Houghton, J. T. & Yihui, D.) 183–237 (Cambridge Univ. Press, 2001).
2. Schimel, D. S. *et al.* *Nature* **414**, 169–172 (2001).
3. Jackson, R. B., Banner, J. L., Jobbágy, E. G., Pockman, W. T. & Wall, D. H. *Nature* **418**, 623–626 (2002).
4. Pacala, S. W. *et al.* *Science* **292**, 2316–2320 (2001).
5. Houghton, R. A. *et al.* *Science* **285**, 574–578 (1999).
6. Van Auken, O. *Annu. Rev. Ecol. Syst.* **31**, 197–215 (2000).

7. Guo, L. B. & Gifford, R. M. *Global Change Biol.* **8**, 345–360 (2002).
8. Sims, P. L. & Singh, J. H. *J. Ecol.* **66**, 573–597 (1978).
9. Jackson, R. B. *et al.* *Proc. Natl Acad. Sci. USA* **94**, 7362–7366 (1997).
10. DeFries, R. S. *et al.* *Int. J. Remote Sensing* **19**, 3141–3168 (1998).

11. Tilman, D. *et al.* *Ecology* **81**, 2680–2685 (2000).
12. Archer, S. *et al.* in *Global Biogeochemical Cycles in the Climate System* (eds Schulze, E.-D. *et al.*) 115–138 (Academic, San Diego, 2001).
13. Schlesinger, W. H. & Pilmanis, A. M. *Biogeochemistry* **42**, 169–187 (1998).

HIV

A tough viral nut to crack

Roger J. Pomerantz

HIV cannot multiply in certain human cells unless it expresses a protein called Vif, the function of which has finally been revealed. It seems that it overcomes a human protein that would otherwise block viral replication.

Nowadays, researchers generally succeed in working out what viral proteins do within a few years of their discovery. This is certainly the case for nearly all proteins from the type 1 human immunodeficiency virus (HIV-1), whose genome is arguably the most extensively studied 9,700 or so bases of genetic sequence on the planet. Nonetheless, one of the HIV-1 proteins has remained an enigma wrapped in a riddle since its discovery in the mid-1980s (refs 1, 2). From the report by Sheehy and colleagues³ on page 646 of this issue, it seems that the mystery has been solved, at least in part.

The protein in question is called Vif, for 'viral infectivity factor': it is an accessory protein in HIV-1 that is also found in all other primate immunodeficiency viruses. Viral accessory proteins are generally not needed for replication and survival; they are neither structural nor regulatory. But they are needed under certain 'stressful' condi-

tions imposed on the virus by its cellular home. Vif seems to be largely unique among HIV-1's accessory proteins. For example, if certain human immune cells — T lymphocytes, monocytes or macrophages, the main reservoirs for HIV-1 *in vivo* — are infected *in vitro* with HIV-1 strains carrying mutant Vif, those strains produce progeny that are, to all intents and purposes, 'dead'⁴. By contrast, mutation of other HIV-1 accessory proteins, such as Nef, Vpr and Vpu, leads to debilitated but still reproducing viruses.

It turns out that HIV-1 with a mutant Vif protein can produce infectious virions — viral particles — from only a few select human cell lines ('producer' cell lines)⁵. When infected with Vif-deficient viruses, certain producer cells, described as 'non-permissive', yield defective progeny viruses that cannot infect target cells and in which reverse transcriptase, a crucial viral enzyme, is largely inactive^{4,6,7}. But infectious, replicating virions can be produced in the absence of

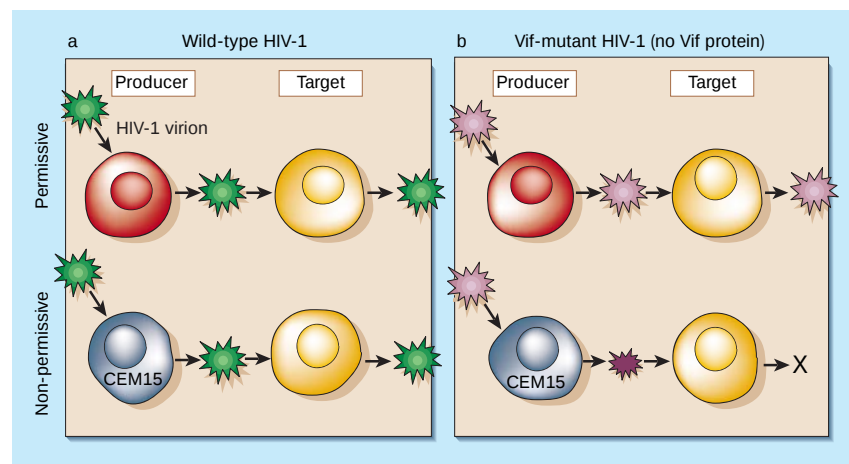


Figure 1 How HIV-1 can overcome our cells' defences. The crucial HIV-1 protein here is Vif. **a**, Normal, infectious HIV-1 can be produced in most human cell types ('producer' cell lines) and can infect and reproduce in most target cells. **b**, HIV-1 particles that lack Vif can produce infectious progeny only in 'permissive' cell types. Non-permissive cells produce non-infectious progeny. Sheehy *et al.*³ have found that non-permissive cell types express the protein CEM15, which therefore presumably blocks the production of new, infectious viral particles from Vif-deficient HIV-1. By extension, the Vif protein must be required to overcome CEM15, explaining why normal HIV-1 can produce infectious viral particles in non-permissive cells (**a**). How Vif suppresses CEM15 is not known.

Vif in 'permissive' cell lines. So the infectivity of Vif-deficient HIV-1 seems to depend on the producer cell type (Fig. 1). Wild-type HIV-1 with an intact Vif gene is not limited in this way.

Over the past decade and longer, numerous laboratories have tried to explain these puzzling observations. Studies have shown^{8–10} that Vif binds to the genomic RNA of HIV-1. But it is unclear how this might lead to the unique characteristics of Vif-deficient viruses. There is clearly some difference between the permissive and non-permissive producer cells, and a breakthrough came in 1998 when two groups^{11,12} proposed that there is a factor, present in some human cells but not in others, that can inhibit the replication of the mutant HIV-1 but is overcome by the Vif protein of the unmutated virus. This cellular factor seemed to alter Vif-deficient HIV-1 during late stages of its life cycle in non-permissive producer cells^{11,12}. Sheehy *et al.*³ now have strong evidence that there is indeed a human protein that inhibits HIV-1 but whose effects are suppressed by Vif.

Sheehy *et al.* looked at two genetically related cell lines, one permissive and one non-permissive. Using a novel combination of molecular-biological approaches, the authors identified a protein — named CEM15, after the cell line in which it was discovered — that was needed for the non-permissive cells to repress the infectivity of the Vif-deficient HIV-1. Sheehy *et al.* have done a superb job of showing that CEM15 is present in all non-permissive cells but not in permissive cells. They also expressed CEM15 in normally permissive cells and found that they became non-permissive to the Vif-deficient HIV-1 (but not to the wild-type virus).

So there is an unarguable correlation between the production of CEM15 and the failure of Vif-deficient HIV-1 particles to produce infectious progeny in non-permissive cells (Fig. 1). Permissive cells, by contrast, do not produce CEM15 and therefore allow the Vif-deficient HIV-1 to replicate freely. All of this suggests that Vif is needed to overcome the human CEM15 protein and allow effective viral replication in non-permissive human cells. The human protein seems to represent one host mechanism — perhaps one of many¹³ — that can inhibit the replication of retroviruses such as HIV-1.

This paper³ takes us a giant leap forward in exploring one of the final frontiers of HIV-1's molecular biology. But it's not yet clear how CEM15 blocks the replication of Vif-deficient HIV-1. Nor is it known whether this protein binds directly to Vif inside human cells infected with wild-type HIV-1, or whether Vif operates in another way. Perhaps one clue lies in CEM15's similarity to two other human proteins: APOBEC-1 (the catalytic subunit of an enzyme that 'edits' messenger RNA¹⁴) and phorbol-1 (a protein induced by phorbol

esters — compounds that stimulate many human cells). All of these proteins have a zinc-coordinating motif, which is important in cytidine deaminase enzymes (including APOBEC-1) in virtually all organisms. It will be interesting to see whether CEM15 is involved in RNA editing, especially as Vif clearly binds to viral genomic RNA^{8–10}.

One implication of Sheehy and colleagues' results is that our cells have a means of silencing HIV-1 in some situations, and at a specific stage of the viral life cycle. This might not be the only cellular virus-silencing tool. For instance, inactive human T lymphocytes expressing the CD4 protein (which HIV-1 uses to latch onto cells) are a key HIV-1 reservoir *in vivo*, yet are difficult to infect with the virus *in vitro*. These cells also have several methods of decreasing the stability and activity of the virus after it has entered host cells but before its genome integrates with that of its host. The results of two recent studies of 'RNA interference' also suggest that this mechanism, which inhibits gene expression, might be used by certain host cells to decrease HIV-1 infection^{15–17}.

Another implication is that the HIV-1 Vif protein and its interactions with CEM15 — whether direct or indirect — provide us with a unique drug target. It might be possible to develop small-molecule drugs that target Vif in the major cell types in which HIV-1 lurks *in vivo*. Another approach could be to inhibit the aggregation of Vif proteins into multimers¹⁸, a process that is required for Vif to function in non-permissive cells. Understanding how CEM15 inhibits HIV-1 replication, and how this inhibition is overcome by Vif, will be crucial in designing drugs to combat this viral accessory protein.

The work of Sheehy *et al.*³ shows that even

the hardest viral nut can be cracked by combining the perseverance of molecular virology laboratories with new molecular techniques. Nevertheless, the story will surely not end here. I predict that this paper will spur the hunt for cellular antiviral molecules and will lead to a heightened understanding of the molecular mechanisms of HIV-1. It also reveals an HIV-1 accessory protein that is a good target for future antiviral drugs — more of which are urgently needed in both developed and developing countries. ■

Roger J. Pomerantz is at the Center for Human Virology, Thomas Jefferson University, 1020 Locust Street, Suite 329, Philadelphia, Pennsylvania 19107, USA.

e-mail: roger.j.pomerantz@mail.tju.edu

1. Fisher, A. G. *et al.* *Science* **237**, 888–893 (1987).
2. Strebel, K. *et al.* *Nature* **328**, 728–773 (1987).
3. Sheehy, A. M., Gaddis, N. C., Choi, J. D. & Malim, M. H. *Nature* **418**, 646–650 (2002); advance online publication, 14 July 2002 (doi:10.1038/nature00939).
4. von Schwedler, U., Song, J., Aiken, C. & Trono, D. *J. Virol.* **67**, 4945–4955 (1993).
5. Gabuzda, D. H. *et al.* *J. Virol.* **66**, 6489–6495 (1992).
6. Goncalves, J., Korin, Y., Zack, J. & Gabuzda, D. *J. Virol.* **70**, 8701–8709 (1996).
7. Dornadula, G., Yang, S., Pomerantz, R. J. & Zhang, H. *J. Virol.* **74**, 2594–2602 (2000).
8. Zhang, H., Pomerantz, R. J., Dornadula, G. & Sun, Y. *J. Virol.* **74**, 8252–8261 (2000).
9. Khan, M. A. *et al.* *J. Virol.* **75**, 7252–7265 (2001).
10. Dettenhofer, M., Cen, S., Carlson, B. A., Kleinum, L. & Yu, X.-F. *J. Virol.* **74**, 8938–8945 (2000).
11. Madani, N. & Kabat, D. *J. Virol.* **72**, 10251–10255 (1998).
12. Simon, J. H. M., Gaddis, N. C., Fouchier, R. A. M. & Malim, M. H. *Nature Med.* **4**, 1397–1400 (1998).
13. Pryciak, P. M. & Varmus, H. E. *J. Virol.* **66**, 5959–5966 (1992).
14. Bhattacharya, S., Navaratnam, N., Morrison, J. R., Scott, J. & Taylor, W. R. *Trends Biochem. Sci.* **19**, 105–106 (1994).
15. Novina, C. D. *et al.* *Nature Med.* **8**, 681–686 (2002).
16. Jacque, J.-M., Triques, K. & Stevenson, M. *Nature* **418**, 435–438 (2002); advance online publication, 26 June 2002 (doi:10.1038/nature00896).
17. Pomerantz, R. J. *Nature Med.* **8**, 659–660 (2002).
18. Yang, S. C., Sun, Y. & Zhang, H. *J. Biol. Chem.* **276**, 4889–4893 (2001).

Obesity

Keeping hunger at bay

Michael W. Schwartz and Gregory J. Morton

Many different hormones control our weight and appetite. The discovery of another hormone, which suppresses appetite for up to 12 hours, may lead to a better understanding of this complex control system.

Spurred on both by a series of remarkable discoveries and by the emergence of obesity as a world health problem, research into how our bodies control our appetite and weight continues at an unparalleled pace. The result is an increasingly clear insight into the workings of a fascinating but complex neuroendocrine system, in which circulating hormones convey information about energy balance (the difference between energy intake and expenditure) to brain pathways that control eating and energy output. The process of discovery has been rather like peeling back the layers of an

onion, as new molecules involved in the system continue to be uncovered — although in some instances it is not the molecule itself but its role in controlling food intake that is discovered. Such is the case for the hormone peptide YY_{3–36} (PYY_{3–36}), as Batterham and colleagues¹ report on page 650 of this issue.

Hormones that regulate food intake can be separated into those that act rapidly to influence individual meals, and those that act more slowly to promote the stability of body fat stores. Long-term regulators include insulin and leptin, which are released into the blood in proportion to the amount

[illegible]

Table 1. *Continued*

and the β parameter is defined as

TABLE 1

[illegible]

melanocortin-producing neurons¹, thereby reducing food intake. The suppression of feeding induced by PYY₃₋₃₆ — like that induced by leptin or insulin — can therefore be predicted to require intact neuronal melanocortin signalling.

These new insights provide a useful framework for understanding how obesity occurs and how it might be treated. But many questions remain to be answered. For instance, does PYY₃₋₃₆ participate in the long-term control of body fat stores, the short-term control of individual meals, or both? Are factors that control its secretion linked to weight changes? At the cellular level, the hypothesis that PYY₃₋₃₆ inhibits NPY/AgRP-expressing neurons through the Y2 receptor implies that signalling involving cyclic AMP is a major determinant of the activity of these neurons (as it is this signal that is blocked by PYY₃₋₃₆). Yet we have no idea what inputs normally drive such signalling in these neurons.

From the clinical perspective, analogues of PYY₃₋₃₆ and other molecules that reduce food consumption by activating the Y2 receptor will no doubt generate interest for their potential in treating obesity. The use of such drugs — perhaps in combination with interventions that reduce signalling by

ghrelin or increase signalling at the neuronal melanocortin, leptin or insulin receptors — may improve obesity treatments to the extent that significant weight loss can be achieved and maintained indefinitely. If we are to combat the global obesity epidemic, such breakthroughs are urgently needed. ■

Michael W. Schwartz and Gregory J. Morton are in the Division of Metabolism, Endocrinology and Nutrition, University of Washington and Harborview Medical Center, 325 9th Avenue, Seattle, Washington 98104-2499, USA.
e-mail: mschwartz@u.washington.edu

1. Batterham, R. L. *et al.* *Nature* **418**, 650–654 (2002).
2. Considine, R. V. *et al.* *J. Clin. Invest.* **95**, 2986–2988 (1995).
3. Bagdade, J. D., Bierman, E. L. & Porte, D. Jr *J. Clin. Invest.* **46**, 1549–1557 (1967).
4. Zhang, Y. *et al.* *Nature* **372**, 425–432 (1994).
5. Woods, S. C., Lotter, E. C., McKay, L. D. & Porte, D. Jr *Nature* **282**, 503–505 (1979).
6. Gibbs, J., Young, R. C. & Smith, G. P. *J. Comp. Physiol. Psychol.* **84**, 488–495 (1973).
7. Cummings, D. E. *et al.* *Diabetes* **50**, 1714–1719 (2001).
8. Stanley, B. G., Kyrkouli, S. E., Lampert, S. & Leibowitz, S. F. *Peptides* **7**, 1189–1192 (1986).
9. Fan, W. *et al.* *Nature* **385**, 165–168 (1997).
10. Sahu, A., Kalra, P. S. & Kalra, S. P. *Peptides* **9**, 83–86 (1988).
11. Schwartz, M. W. *et al.* *Diabetes* **46**, 2119–2123 (1997).
12. Hahn, T. M., Breininger, J. F., Baskin, D. G. & Schwartz, M. W. *Nature Neurosci.* **1**, 271–272 (1998).
13. Nakazato, M. *et al.* *Nature* **409**, 194–198 (2001).
14. Gobbi, M., Mennini, T. & Vezzani, A. *J. Neurochem.* **72**, 1663–1670 (1999).
15. Cowley, M. A. *Nature* **411**, 480–484 (2001).

Semiconductor physics

One at a time, please

Manfred Bayer

Semiconductor quantum dots could become the basis of the much-talked-about quantum computer. A single-electron 'turnstile' device is a promising way to read out the information being processed.

Physicists are actively seeking to transfer the weird phenomena of the quantum world from their laboratories to the real world. In particular, the prospect of quantum-information processing has attracted considerable attention for its potential to improve the speed and reliability of data handling¹. Information would be encoded in 'quantum bits', and the search is on for a physical system that could form a reliable, controllable quantum bit: on page 612 of this issue, Zrenner *et al.*² report their progress with a candidate from solid-state physics, the 'exciton'.

In contrast to classical bits, which can be in either state 0 or state 1, quantum bits exist as a combination (a linear superposition) of two quantum logic states, represented as $|0\rangle$ and $|1\rangle$. In a quantum computer, the quantum bits first have to be controlled individually in order to initialize the quantum register in which information is stored. Then a controllable interaction between the quantum bits must be established so that the quantum states become entangled. It is this 'entangle-

ment' that is the key to a quantum computer's power: in effect, a rigid coupling is introduced between the quantum bits, which can then no longer be considered individually but are affected simultaneously by a calculational operation. It is thanks to this capacity for parallel processing that a quantum computer should be able to perform calculations much faster than a classical computer.

The original proposals for quantum computers were based on atomic systems¹, such as atoms held in traps, where the quantum bit is formed by two energy levels between which an atomic electron can make transitions. Now that semiconductor quantum dots have been synthesized, it opens up the possibility of mimicking these approaches in a solid-state environment. Quantum dots, tiny clusters of semiconductor material, are often called 'artificial atoms', because the charge carriers in these systems (electrons or holes) can only occupy a restricted set of energy levels, just like the electrons in an atom. In fact, quantum dots offer a variety of two-level systems, based on charge or spin

(or both). One such two-level system is a coupled electron-hole pair — an exciton. The absence (equivalent to the state $|0\rangle$) and presence (state $|1\rangle$) of an exciton in a semiconductor quantum dot could represent the two levels of a quantum bit.

But to work as a quantum bit, it is essential that the quantum states of this exciton can be prepared, controlled and measured with high accuracy. Zrenner *et al.*² have taken a remarkable step forward by showing that coherent manipulation of an exciton quantum bit is possible by applying short laser pulses to a single quantum dot. Further, they demonstrate that the quantum-bit state can be measured with extremely high efficiency, through its transformation into an electric current.

Optical excitation using a picosecond laser pulse creates an exciton in the quantum dot, as an electron is lifted from the valence band into the conduction band of the semiconductor: the electron is coupled to the 'hole' it left behind in the valence band, forming an exciton. In a simplified model, the exciton can be considered as a harmonic oscillator — like the membrane of a loudspeaker — that is driven by the laser field. On the short timescales of the laser pulse, damping of the oscillator does not occur. As the laser intensity increases, equivalent to pushing the loudspeaker membrane harder, the amplitude of the oscillator increases to a maximum value, then decreases. Increasing the laser intensity further sets up a periodic repetition of this excitation and de-excitation process — the creation and annihilation of excitons in the quantum dot, an effect called Rabi oscillation.

Exciton Rabi oscillations in single quantum dots are a recently demonstrated phenomenon^{3–5}, relying on optical detections. A photon may be released as the electron and hole recombine and the exciton is destroyed. But this kind of measurement is rather inefficient, and so photon detection would be of limited use as a measurement protocol for quantum-information processing. However, Zrenner *et al.* have developed a more effective method for detecting the quantum state that involves mapping the quantum state onto a current. Whereas photons can easily just get lost en route to a detector, this current can be measured with high precision because charge must be conserved in a circuit. The quantum dot is placed between two electrodes to which voltage is applied, and the electron and the hole of the exciton tunnel out of the dot, generating a current that reflects the Rabi oscillations taking place.

The situation becomes especially interesting when the laser intensity is such that the exciton oscillator reaches its maximum amplitude. This corresponds to 'full inversion': the prepared quantum state is wholly $|1\rangle$, with no state $|0\rangle$ mixed in. Each laser pulse creates exactly one exciton, subsequently causing a single elementary charge to flow through the detection circuit. Thus,



100 YEARS AGO

The San Francisco correspondent of the *Daily Mail* reports that the people of Santa Barbara, a county of southern California, are terror-stricken owing to the increasing frequency and severity of the earthquake shocks, of which there were seventy-five from July 27–31. The most destructive was that at the town of Los Alamos, at 1.20 a.m. on July 31. All the brick buildings were thrown to the ground, but the frame buildings generally escaped serious injury except to their windows... The shock lasted thirty seconds, and seems to have had a spiral motion. Goods were hurled from the shelves of the stores and piled in the middle of the rooms; even heavy desks were tossed about. The inhabitants ran into the streets in a panic, for in the morning between 7.25 and 7.30 there were additional shocks, and just before nine two more. ... The earth continues to tremble at intervals, and the countryside is said to be changing appearance. From *Nature* 7 August 1902.

50 YEARS AGO

Dr. Conant points out that we are now dealing with the consequences of a new social phenomenon. Until the advent of nuclear weapons, developments in defensive or offensive weapons depended on the special application of the publicly known facts and principles of physics and chemistry. Any new scientific knowledge involved had no revolutionary consequences for the advance of science. The flowering of the whole field of nuclear physics, on the contrary, depended on the expenditure of a large sum of public money, justified in the first instance in terms of the destructive power of a weapon required in a desperate global struggle. Inevitably nuclear physics, thus born of a war-time necessity, was marked from its birth with the secrecy which was anathema to nineteenth-century men of science. The advance of one whole area of science was thus equated with national security; and Dr. Conant insists that the assumption that science knows no frontier is automatically cancelled and we must face the consequences. We must distinguish sharply between the special position of nuclear science and that of the other fields of physics and chemistry. This is the first step towards preventing the spread of undue secrecy and thereby assisting the progress of science and promoting the welfare of the international community of science. From *Nature* 9 August 1952.

the temporal sequence of equidistantly spaced laser pulses is transformed into a sequence of unit-charge current pulses. For a laser-pulse repetition rate of 82 MHz, the expected current is 13.1 pA, which agrees well with the value observed by Zrenner *et al.* of 12.6 pA. The experiment is effectively a single-electron turnstile device — one electron passes per operation cycle in a quasi-deterministic fashion.

But several problems must be overcome if this is to be the basis of viable quantum-computing technology. To give just two examples, Zrenner *et al.* found that, as the excitation power is increased, the oscillations become considerably damped, seeming to die away. The origin of this damping and a strategy for overcoming it need to be worked out. Second, it must be shown that

coherent processing occurs with subsequent laser pulses of the exciton quantum state generated by the first laser pulse.

Nevertheless, looking to the future of coherent optoelectronics, the work of Zrenner *et al.* will no doubt prove significant. Their single-electron turnstile may become one of the building-blocks of a new generation of optoelectronic devices of unprecedented performance. ■

Manfred Bayer is in the Department of Experimental Physics II, Universität Dortmund, D-44221 Dortmund, Germany.

e-mail: manfred.bayer@physik.uni-dortmund.de

1. Bouwmeester, D., Ekert, A. & Zeilinger, A. (eds) *The Physics of Quantum Information* (Springer, Berlin, 2000).
2. Zrenner, A. *et al.* *Nature* **418**, 612–614 (2002).
3. Stievater, T. H. *et al.* *Phys. Rev. Lett.* **87**, 133603 (2000).
4. Kamada, H. *et al.* *Phys. Rev. Lett.* **87**, 246401 (2001).
5. Htoon, H. *et al.* *Phys. Rev. Lett.* **88**, 087401 (2002).

Structural biology

Calcium callisthenics

N. Michael Green and David H. MacLennan

The Ca^{2+} -ATPase is one of the best characterized of the pumps that move ions across cell membranes. The latest snapshot of the pump in another of its many conformations is a major step in understanding its dynamics.

The action of skeletal muscle depends on the flow of calcium ions into and out of a storage compartment — the sarcoplasmic reticulum — in muscle cells. Release of Ca^{2+} into the cytosol stimulates muscle contraction. The reverse flow to replenish the Ca^{2+} store and permit muscle relaxation is controlled by a membrane protein, the Ca^{2+} -ATPase, which occurs in other cells but in muscle cells constitutes up to 90% of the membrane protein. In their paper on page 605 of this issue¹, Toyoshima and Nomura provide a spectacular view of the conformational changes that take place between this pump in its calcium-free and calcium-bound states.

Ca^{2+} -ATPase is part of a large family of enzymes, all of which contain a specific amino-acid residue, aspartic acid, that is phosphorylated by ATP only in the presence of the appropriate cation. The cation concerned enters the pump from the cytoplasmic side, is trapped within the membrane and is released on the luminal or extracellular side through gates that are controlled by conformational changes in the enzyme. The kinetic cycle for the Ca^{2+} -ATPase is shown in Fig. 1. It consists of a complex, reversible sequence of phosphorylation and dephosphorylation.

Two years ago, Toyoshima and colleagues published² a crystal structure of the muscle Ca^{2+} -ATPase, SERCA1a. This both confirmed widely held concepts³ about the mechanism of a Ca^{2+} pump, and provided fresh insights into its action. This structure

showed the Ca^{2+} -bound form, E1Ca₂ (Fig. 2b), with its three cytoplasmic domains well separated, but joined to one another or to membrane helices by extended links (this contrasts with the compact structure of these domains that was observed in an earlier, lower-resolution structure for E2P; ref. 4). Only the catalytic P domain was anchored more firmly, by substantial extensions of the transmembrane α -helices M4 and M5. Calcium was bound in the transmembrane domain to two high-affinity sites formed by the precise juxtaposition of M4, M5, M6 and M8, with the 12–14 ligands to the Ca^{2+} ions linking these helices together.

To complete the cycle and release Ca^{2+} to the lumen of the sarcoplasmic reticulum, ATP is bound in the jaw formed by the N and P domains, which closes, allowing phosphorylation of the crucial aspartic acid (D351). A slow rearrangement of phosphoenzyme

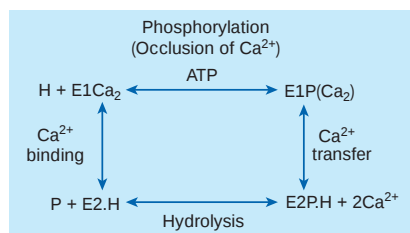


Figure 1 The basic reaction cycle of Ca^{2+} -ATPase. Binding of Ca^{2+} and phosphorylation alternate with transfer of Ca^{2+} and hydrolysis. The structural transformation discussed in the text is that between E2H and E1Ca₂.

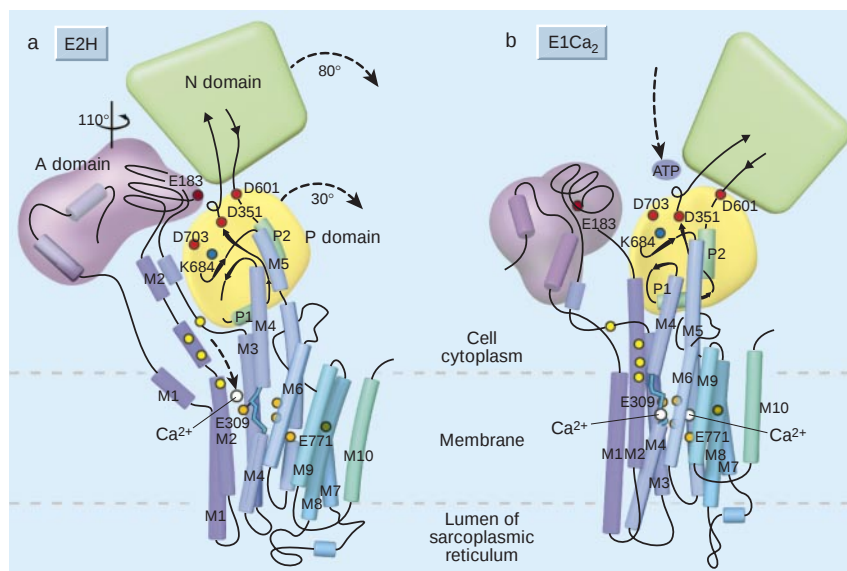


Figure 2 Structure and function in a Ca^{2+} pump. Shown here are the Ca^{2+} -free (E2H) and Ca^{2+} -bound (E1 Ca_2) states of SERCA1a, illustrating the effects of Ca^{2+} binding on the pump structure. Three domains, P (phosphorylation), N (nucleotide-binding) and A (actuator), lie in the cytosol. The P domain, with N inserted, connects the tops of transmembrane helices M4 and M5. Two β -strands, which continue from these helices, form additional anchors that are part of the central β -sheet. This domain includes a cluster of critical catalytic residues: the phosphorylation site D (aspartic acid) 351, D703 and D601, and K (lysine) 684. Some of the Ca^{2+} ligands (orange) on transmembrane helices M4, M5, M6 and M8, and some of the polar or negatively charged residues (yellow) that line a possible Ca^{2+} entry pathway, are shown in relation to bound or unbound Ca^{2+} (white). a, The E2H state, as determined in the new work¹. Entry to the Ca^{2+} sites is blocked by E (glutamic acid) 309 in M4. The P, N and A domains form a compact cluster. The A domain is rotated to meet the P domain and carries E183 into the catalytic site, where it may be involved in hydrolysis of E2P. b, The E1 Ca_2 state. In the crystal structure², the cytoplasmic domains have moved apart (they are known to be mobile in solution). ATP can bind in the cleft between N and A, but because it induces a new conformation its location is only partly defined. (a, Based on refs 1, 2; b, based on ref. 2.)

E1P(Ca_2) disrupts the Ca^{2+} -binding sites, opening the luminal gate. This allows water to enter the catalytic site and hydrolyse the phosphoenzyme E2P, generating the E2H conformation. A channel then opens to allow cytoplasmic Ca^{2+} access to the binding sites to re-form E1 Ca_2 .

In the new paper¹, Toyoshima and Nomura describe how they have captured SERCA1a at 3.1 Å resolution in the E2 configuration (Fig. 2a). The pump is protonated⁵, and there is good evidence that protons are counter-transported⁶, like potassium ions in another well-characterized ion pump, the Na^+/K^+ pump. The protonated residues are not defined, but four carboxyl groups that provide Ca^{2+} ligands are likely candidates. In comparison with E1 Ca_2 , there are complex changes in the transmembrane Ca^{2+} -binding site and the 'stalk' connections, which are amplified even further by movement of the cytoplasmic domains, which gather to form a compact cluster (Fig. 2a).

In the dephosphorylated segment of the cycle ($\text{E2H} \leftrightarrow \text{E1} \leftrightarrow \text{E1Ca} \leftrightarrow \text{E1Ca} \leftrightarrow \text{E1Ca}_2$), the pump loses protons, binds one Ca^{2+} , undergoes a conformational change, and then cooperatively binds a second Ca^{2+} . The structural correlates of this transition are

illustrated in Fig. 2. Large-scale movements of the N, P and A domains occur with little internal change. They are coupled to changes in tilt and position of helices M1–M6, but M7–M10 are almost unmoved. The bottom section of M5, below glutamic acid (Glu) 771, remains fixed, packing against M7. But the middle section, with its cytoplasmic extension bonded to the L67 loop, tilts back, straightening the helix and carrying the top section—an integral part of the P domain—through a 30° rotation. The largest movements of Ca^{2+} ligands are in M6, which becomes more helical on binding Ca^{2+} , several ligands rotating through 90°. The M4 helix moves down by 5 Å, carried by the tilting of M5. The net result is that the Ca^{2+} ligands, which are dispersed in E2H, come together to form compact sites in E1 Ca_2 .

These changes stabilize a new configuration of the surrounding helices, and the effects are transmitted to the P domain, to M1, M2, M3, and to the A domain. Parts of the structure are likely to be thermally labile, the molecule being stabilized in E2H by hydrogen bonds between the protein domains and by nonpolar interactions (see Fig. 1, right, on page 606), and in E1 Ca_2 by multiple ligands to Ca^{2+} . In the cytoplasm,

binding of Ca^{2+} leads to an open domain structure, which can rearrange further after phosphorylation. The coordinating role of the P domain is clear from its many interactions with N, A and the transmembrane and stalk domains.

The E2H structure highlights the access routes for Ca^{2+} . An early model⁷ of the trans-formation between E1 and E2 proposed that Ca^{2+} was bound to exposed luminal sites in E2 or to cytoplasmic sites in E1. Lack of competition between these sites suggested⁸ that access to the lumen was confined to stage E2P. Consistently, in E2H, closure of luminal loops L34 and L78 blocks access from this side. On the cytoplasmic side, Glu 309 in M4 rotates to lie at the bottom of a negatively charged, aqueous pocket in E2H (ref. 9), blocking access to the main Ca^{2+} -binding sites. A small movement of Glu 309, possibly triggered by ionization, might open this entrance and allow Ca^{2+} to enter the binding sites. This would be consistent with a gating role, proposed to explain the properties of mutants of Glu 309 in the Ca^{2+} pump and Glu 327 in the Na^+/K^+ pump¹⁰.

Further progress will, ideally, require determination of the structures for the ATP-bound (E1ATP Ca_2) and phosphorylated (E1P(Ca_2) and E2PH) states. However, these intermediate conformations are labile, and stabilizing ligands, such as thapsigargin in the E2H structure, will be needed, making them structural guides rather than true intermediates. Although it may be possible to define further conformational steps between binding of calcium and phosphorylation, it may be more realistic to represent the process as a balance between global conformations, tipped one way or the other by ligands or by phosphorylation. Toyoshima and Nomura¹ have illuminated one small step in the cycle of the Ca^{2+} pump. But their depiction of the dynamic interrelationships among the key residues and interlinked domains is a giant step forward in understanding the mechanism of the ATPase family that includes SERCA1a.

N. Michael Green is at the National Institute of Medical Research, Mill Hill, London NW7 1AA, UK. e-mail: mgreen@nimr.mrc.ac.uk

David H. MacLennan is in the Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario M5G 1L6, Canada. e-mail: david.maclennan@utoronto.ca

- Toyoshima, C. & Nomura, H. *Nature* **418**, 605–611 (2002).
- Toyoshima, C., Nakasako, M., Nomura, H. & Ogawa, H. *Nature* **405**, 647–655 (2000).
- MacLennan, D. H., Rice, W. J. & Green, N. M. *J. Biol. Chem.* **272**, 28815–28818 (1997).
- Zhang, P. et al. *Nature* **392**, 835–839 (1998).
- Mintz, E. & Guillain, F. *Biochim. Biophys. Acta* **1318**, 52–70 (1997).
- Levy, D. et al. *J. Biol. Chem.* **265**, 19524–19534 (1990).
- de Meis, L. & Vianna, A. L. *Annu. Rev. Biochem.* **48**, 275–292 (1979).
- Myung, J. & Jencks, W. P. *FEBS Lett.* **278**, 35–37 (1991).
- Clarke, D. M. et al. *J. Biol. Chem.* **264**, 11246–11251 (1989).
- Vilsen, B. & Andersen, J. P. *Biochemistry* **37**, 10961–10971 (1998).

Obituary

W. Maxwell Cowan (1931–2002)

The remarkable successes of modern biomedical research stem from a combination of factors: talented scientists have enjoyed an intellectual climate that encourages innovation and excellence, and they have received growing funding support from public and private sources. W. Maxwell Cowan, who died on 30 June, will be remembered for his outstanding contributions relating to each of these factors. He was a renowned neuroanatomist who charted fundamental aspects of brain circuitry and development. He helped to shape modern neuroscience as an integrated yet interdisciplinary field, not least through his editorship of several journals. In the last decade of his working life, he moved onto a broader stage at the Howard Hughes Medical Institute (HHMI), where he orchestrated the expansion of its support for research.

For more than a century, neuroanatomists have worked to decipher the fabulously complex neural circuitry that underlies brain function. Progress has depended on a succession of methodological advances for analysing axonal pathways with increasing precision and sensitivity. Cowan entered this field on moving in 1953 to Oxford, UK, from South Africa, where he had studied under Raymond Dart, the discoverer of the early prehuman fossil *Australopithecus africanus*.

At Oxford, Cowan carried out doctoral research under Wilfrid LeGros Clark, a leading neuroanatomist. Cowan and his colleagues (most notably T. P. S. Powell) used the then new silver-staining methods based on axonal degeneration to chart the connections of many different brain structures, particularly those involving the thalamus, hypothalamus and hippocampus. Although the 'degeneration' methods had serious technical limitations, these observations formed the basis for much of our current understanding of brain circuitry.

In 1966, Cowan moved to the University of Wisconsin and, two years later, to St Louis, to head the Department of Anatomy at Washington University School of Medicine — where we now work. Around this time, discoveries in other laboratories revealed that neurons have a specialized transport system for conveying materials between their cell bodies and synaptic terminals. Cowan recognized that if tracer molecules were piggybacked onto this system, many of the drawbacks of



A founder of modern neuroscience, and a guiding hand in biomedical research

degeneration-based methods for tracing neuronal pathways could be circumvented. He and his colleagues introduced a method for injecting radioactive amino acids into specific sites in the brain, to reveal the distribution of labelled proteins that underwent anterograde transport (that is, from neuronal cell bodies to their axonal terminations). This was a key step in launching modern neuroanatomy, which now uses a variety of tracers — transported in an anterograde, retrograde or bidirectional manner — to reveal how neurons connect with one another. Increasingly widespread application of these methods at Washington University and elsewhere generated a revolution in neuroanatomy in the 1970s.

An advantage of the tracer-based method was its applicability to the developing brain, which is notoriously difficult to study with the classical degeneration methods. Over the next 20 years, much of Cowan's research was directed at questions of how connections form and are refined during development, and how they can be modified in the adult. He also explored factors that influence neuronal survival, inspired by his long friendship with Viktor Hamburger, whose own studies of cell death during development led to the identification of nerve growth factor and other agents that control the growth of neurons.

In the late 1960s, an identity crisis was brewing in brain research — the number of neuroscientists was increasing rapidly, yet they were typically scattered across

departments of anatomy, physiology, biology and pharmacology. Neuroscience did not exist as a generally recognized field of investigation or of graduate training. Over the ensuing two decades, Cowan was prominent in helping to launch and define neuroscience as a broad yet integrated discipline. He took over editorship of the faltering *Journal of Comparative Neurology* in 1969, reviving it as a major neuroscience journal. He attracted a diverse group of scientists to his department, renamed it the Department of Anatomy and Neurobiology, and led it to pre-eminence as a centre of interdisciplinary neuroscience research. An infamous feature that he instituted was the Saturday morning seminar series, where full attendance was expected.

In 1980, Cowan moved to the Salk Institute, where he served as vice-president from 1982 to 1986. The talented young investigators in his laboratory continued to produce striking results, but increasingly his time was occupied with administration. He became the inaugural editor of *The Journal of Neuroscience* in 1982 and led it to prominence. His recruitment, in 1984, to the Medical Advisory Board of the HHMI gave him a wider arena in which to exercise his scientific and organizational talents to their fullest effect.

The HHMI is the largest private supporter of biomedical research in the United States, and his advice helped bring neuroscience into the portfolio of programmes sponsored by the institute. In 1988, he became vice-president and chief scientific officer of the HHMI, and over the subsequent 12 years played a pivotal role in shaping the institute's scientific agenda. During this period, the number of HHMI-supported investigators more than doubled, as did the number of academic institutions at which they are located.

Cowan's depth of knowledge of neuroscience was legendary. Equally impressive was his voracious appetite for learning about the full breadth of research activities supported by the HHMI. Cowan was a masterly communicator, equally effective at holding the attention of scientific experts and a lay audience. He has left an indelible stamp on the field of modern neuroscience and on the broader realm of biomedical research. David C. Van Essen and Joseph L. Price
David C. Van Essen and Joseph L. Price are in the Department of Anatomy and Neurobiology, Washington University School of Medicine, 660 S. Euclid, St Louis, Missouri 63110-1093, USA. e-mail: vanessen@v1.wustl.edu

Contrails reduce daily temperature range

A brief interval when the skies were clear of jets unmasked an effect on climate.

The potential of condensation trails (contrails) from jet aircraft to affect regional-scale surface temperatures has been debated for years^{1–3}, but was difficult to verify until an opportunity arose as a result of the three-day grounding of all commercial aircraft in the United States in the aftermath of the terrorist attacks on 11 September 2001. Here we show that there was an anomalous increase in the average diurnal temperature range (that is, the difference between the daytime maximum and night-time minimum temperatures) for the period 11–14 September 2001. Because persisting contrails can reduce the transfer of both incoming solar and outgoing infrared radiation^{4,5} and so reduce the daily temperature range, we attribute at least a portion of this anomaly to the absence of contrails over this period.

We analysed maximum and minimum temperature data⁶ from about 4,000 weather stations throughout the conterminous United States (the 48 states not including Alaska and Hawaii) for the period 1971–2000, and compared these to the conditions that prevailed during the three-day aircraft-grounding period. All sites were inspected for data quality and adjusted for the time of observation.

Because the grounding period commenced after the minimum temperatures had been reached on the morning of 11 September and ended before maximum temperatures were attained on 14 September (at noon, Eastern Standard Time), we staggered the calculation of the average diurnal temperature range (DTR) across adjacent days (for example, 11 September maxima minus 12 September minima). We repeated this procedure for the three-day periods immediately before and after the grounding period, and also for the same periods (8–11, 11–14 and 14–17 September) for each year from 1971 to 2000.

DTRs for 11–14 September 2001 measured at stations across the United States show an increase of about 1.1 °C over normal 1971–2000 values (Fig. 1). This is in contrast to the adjacent three-day periods, when DTR values were near or below the mean (Fig. 1). DTR departures for the grounding period are, on average, 1.8 °C greater than DTR departures for the two adjacent three-day periods.

This increase in DTR is larger than any during the 11–14 September period for the previous 30 years, and is the only increase greater than 2 standard deviations away from the mean DTR (s.d., 0.85 °C). Moreover, the 11–14 September increase in DTR

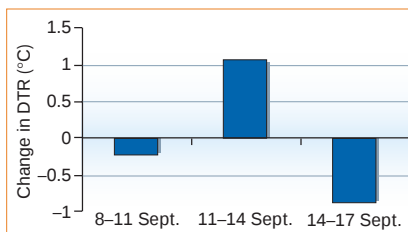


Figure 1 Departure of average diurnal temperature ranges (DTRs) from the normal values derived from 1971–2000 climatology data for the indicated three-day periods in September 2001. These periods included the three days before the terrorist attacks of 11 September; the three days immediately afterwards, when aircraft were grounded and there were therefore no contrails; and the subsequent three days.

was more than twice the national average for regions of the United States where contrail coverage has previously been reported to be most abundant (such as the midwest, northeast and northwest regions)⁷.

Day-to-day changes in synoptic atmospheric conditions can affect regional DTRs⁸. In particular, a lack of cloud cover helps to increase the maximum (and reduce the minimum) temperature. Maps of the daily average outgoing long-wave radiation (OLR)^{9,10} — a proxy for optically thick clouds — show reduced cloudiness (that is, larger OLR) over the eastern half of the United States on 11 September, but more cloud (smaller OLR) over parts of the west. Cloud cover subsequently decreased in the west and increased over much of the eastern half of the country during the next two days, producing predominantly negative three-day OLR changes in the east and positive values in parts of the west.

Our findings indicate that the diurnal temperature range averaged across the United States was increased during the aircraft-grounding period, despite large variations in the amount of cloud associated with mobile

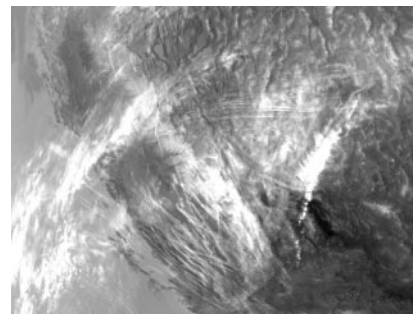


Figure 2 Flight lines: jet contrails can clearly be seen as thin streaks in this satellite image of the southwestern United States.

weather systems (Fig. 2). We argue that the absence of contrails was responsible for the difference between a period of above-normal but unremarkable DTR and the anomalous conditions that were recorded.

David J. Travis*, **Andrew M. Carleton†**, **Ryan G. Lauritsen***

**Department of Geography and Geology, University of Wisconsin–Whitewater, Whitewater, Wisconsin 53190, USA*

e-mail: travisd@uww.edu

†Department of Geography, Pennsylvania State University, University Park, Pennsylvania 16801, USA

1. Changnon, S. A. *J. Appl. Meteorol.* **20**, 496–508 (1981).
2. Travis, D. J. & Changnon, S. A. *J. Weather Modification* **29**, 74–83 (1997).
3. Sassen, K. *Bull. Am. Meteorol. Soc.* **78**, 1885–1903 (1997).
4. Duda, D. P., Minnis, P. & Nguyen, L. *J. Geophys. Res.* **106**, 4927–4937 (2001).
5. Meerkotter, R. *et al. Ann. Geophys.* **17**, 1080–1094 (1999).
6. National Oceanic and Atmospheric Administration. *TD3200/3210 Data Set for 1971–2001* (Natl Climate Data Center, Asheville, North Carolina, 2001).
7. DeGrand, J. Q., Carleton, A. M., Travis, D. J. & Lamb, P. J. *J. Appl. Meteorol.* **39**, 1434–1459 (2000).
8. Karl, T. R. *et al. Bull. Am. Meteorol. Soc.* **74**, 1007–1023 (1993).
9. Liebmann, B. & Smith, C. A. *Bull. Am. Meteorol. Soc.* **77**, 1275–1277 (1996).
10. <http://www.cdc.noaa.gov> (NOAA-CIRES Climate Diagnostics Center, Boulder, Colorado, USA).

Competing financial interests: declared none.

Animal behaviour

Male parenting of New Guinea froglets

Male parental care is exceptionally rare in nature, although one of the most fascinating aspects of New Guinea's biodiversity is the evolution of male care in the frog family Microhylidae¹. Here I report a new mode of parental care: transport of froglets by the male parent, which was recently discovered in two species of microhylid

frogs from the mountains of Papua New Guinea. As the offspring jump off at different points, they may benefit from reduced competition for food, lower predation pressure and fewer opportunities for inbreeding between froglets, which may explain why this unusual form of parental care evolved.

I quantified the parental care behaviour of several species of microhylid frog at the Crater Mountain Biological Research Station, Chimbu Province, Papua New Guinea (6° 43' S, 145° 05' E), which is located on the largest tropical island in



Figure 1 A male *Liophryne schluginhaufeni* transporting froglets. Eleven froglets are visible (out of 22 in total) on the lateral and dorsal surfaces of the adult. The froglet located on top of the father's head is preparing to jump off.

remote rainforest at 800–1,350 m, where the topography and rainfall (6.4 m per year)² are extreme. All microhylid frogs on New Guinea are thought to develop directly from eggs, skipping the aquatic tadpole stage and hatching as miniature versions of the adults³.

Among the many (more than 20) microhylid species at this site, I discovered that, in addition to guarding eggs, males of at least two species (*Liophryne schluginhaufeni* and *Sphenophryne cornuta*) transport the froglets after they have hatched (Fig. 1). I observed 23 froglet-transport events: 9 in *S. cornuta* and 14 in *L. schluginhaufeni*. In all of the 19 cases for which the sex of the transporting individual was ascertained, froglets were transported by the male.

Five entire transport-clutches (from eggs to independent froglets) were seen in *L. schluginhaufeni*, with males carrying froglets for three to nine nights (mean $\pm$ s.d., 6.6 ± 2.6 ; $n=5$) and travelling 0–17 metres each night (7.6 ± 4.19 ; $n=29$) through herbaceous vegetation, and seeking refuge under the leaf litter of the forest floor during the day. Total transportation distances ranged from 34 to 55 m (44.4 ± 8.7 m; $n=5$).

The regular dispersal pattern of froglets has potential benefits. Froglets distributed themselves evenly over time (fewer than seven froglets dispersing per night; 3.6 ± 1.7 ; $n=49$) and space (between 0.4 and 5.5 m; 3.3 ± 0.7 m; $n=49$) by individually jumping off the transporting male. The rewards accrued by the froglets might include less competition for food, a lower chance of predation and a decreased potential for inbreeding because they are widely dispersed.

Male parental care in other frog families includes both attendance and

transport of eggs and tadpoles, but not froglet transport⁴. Froglet transport has only been reported for females in a single species, the Jamaican cave-breeding frog *Eleutherodactylus cundalli*⁵, and has either not been observed in New Guinean frogs⁶ or has been only briefly described without identification of the species or sex of the caring adult⁷.

Uniparental male care is extremely rare in terrestrial vertebrates⁸. Even in groups with male care, most species display maternal or biparental care, or no care at all. The microhylid frogs of New Guinea are the only known large group (over 150 species in about 20 genera) of terrestrial vertebrates in which male care predominates. Comparative cost–benefit analyses should provide insights into the evolution of this behaviour and the role of parental care in the radiation of microhylid frogs on New Guinea, adding to our understanding of the environmental and/or historical conditions under which male parental care evolves.

David Bickford

University of Miami, PO Box 249118, Coral Gables, Florida 33124-0421, USA, and
Wildlife Conservation Society, Papua New Guinea Program, PO Box 277, Goroka, EHP,
Papua New Guinea
e-mail: bickford@bio.miami.edu

1. Bickford, D. *The Ecology and Evolution of Parental Care in the Microhylid Frogs of New Guinea*. Thesis, Univ. Miami (2001).
2. Wright, D. D., Jessen, J. H., Burke, P. & de Silva Garza, H. G. *Biotropica* **29**, 250–260 (1997).
3. Zweifel, R. G. *Bull. Am. Mus. Nat. Hist.* **148**, 411–546 (1972).
4. Crump, M. L. *Adv. Study Behav.* **25**, 109–144 (1996).
5. Diesel, R., Baurle, G. & Vogel, P. *Copeia* **1995**, 354–360 (1995).
6. Simon, M. *Behav. Ecol. Soc.* **14**, 61–67 (1983).
7. Cogger, H. G. in *Reptiles and Amphibians of Australia* 5th edn, 159 (Reed, Sydney, 1996).
8. Clutton-Brock, T. H. *The Evolution of Parental Care* (Princeton Univ. Press, Princeton, New Jersey, 1991).

Competing financial interests: declared none.

Cosmology

Black holes constrain varying constants

There is evidence to suggest that the fine-structure constant, α — a measure of the strength of the electromagnetic interaction between photons and electrons — is slowly increasing over cosmological timescales^{1–4}. As $\alpha = e^2/\hbar c$ (where e is the electronic charge, $\hbar$ is Planck's constant and c is the speed of light), this would call into question which of these fundamental quantities are truly constant. Here we consider black-hole thermodynamics as a test of which constants might actually be variable, discounting those that could lead to a violation of the generalized second law of thermodynamics.

Observational evidence suggests that there has been a variation of $\Delta\alpha/\alpha = -0.72 \pm 0.18 \times 10^{-5}$ over the past 6–10 billion years¹. This result could be interpreted as supporting some non-standard cosmological theories that invoke varying the speed of light^{5–7} or the electronic charge⁸. It has been shown⁹ that a varying- c cosmology, through changes to standard units, can be rephrased as a varying- e theory, similar to the one proposed earlier⁸. If attention is restricted to electromagnetic phenomena, there is no observational difference between the theories, and either c or e could account equally well for the variation in α . However, there may be fundamental theoretical reasons concerned with gravitation to favour varying c over varying e .

One way to introduce gravitation into the discussion is through the theory of black-hole thermodynamics. Entropy is associated with the area of a black hole's event horizon^{10,11}, leading to a generalized second law of thermodynamics in which the event horizon's area may only decrease if there is a corresponding increase in the conventional entropy of the black hole's environment^{10,11}.

In the case of a non-rotating black hole with electric charge Q and mass M , the area of its event horizon, A_H , is obtained in conventional general-relativity theory from the Reissner–Nordström solution of Einstein's field equations

$$A_H = 4\pi r^2 \quad (1)$$

where

$$r = \frac{G}{c^2} [M + \sqrt{M^2 - Q^2/G}] \quad (2)$$

The entropy of a black hole is given by $S_H = (kc^3/G\hbar)A_H/4$, which becomes

$$S_H = \frac{k\pi G}{\hbar c} [M + \sqrt{M^2 - Q^2/G}]^2 \quad (3)$$

where k is Boltzmann's constant.

An increase in the magnitude of the electric charge, Q , with c and G remaining constant, implies a reduction in the area of the event horizon. By contrast, a decrease in the speed of light, c , would lead to an increase in event-horizon area. Thus the two contending alternatives for an increase in α produce opposite outcomes as far as black-hole entropy is concerned.

It could be argued that a reduction in event-horizon area implies a violation of the generalized second law of thermodynamics, and so the fundamental electric charge therefore cannot increase. However, before we can be secure in that interpretation, several conditions must be satisfied. The black hole will radiate heat into its environment through the Hawking process, and, as Q changes, the temperature will also change. For the second law of thermodynamics to be violated, the black hole must not raise the entropy of the environment by more than its own entropy decreases. This condition is readily satisfied by immersing the black hole in a heat bath of equal temperature and allowing the heat radiation to change isentropically as the charge varies.

Furthermore, equation (3) is based on standard gravitational theory. In a non-standard theory that involves varying e or c , the formula for the area of the event horizon might differ¹². Also, the Hawking process may be modified in a way that alters the relationship between temperature, entropy and event-horizon area. Equation (3) must then be considered as an approximation of the limit of small variation of 'constants'. However, it is unlikely that minor modification of equation (3) will reverse the sign of the relationship between charge and event-horizon area.

Moreover, in the standard theory there is a maximal electric charge, given by $Q^2 = M^2$, above which the event horizon disappears and the black hole is replaced by a naked singularity. A modified theory might alter the value of this maximal charge, but there will still be a limit above which any increase in charge will create a naked singularity, violating the cosmic-censorship hypothesis¹³.

Our arguments, although only suggestive, indicate that theories in which e increases with time are at risk of violating both the second law of thermodynamics and the cosmic-censorship hypothesis. Thus, black-hole thermodynamics may provide a stringent criterion against which contending theories for varying 'constants' should be tested.

P. C. W. Davies*, Tamara M. Davis†, Charles H. Lineweaver†

*Australian Centre for Astrobiology, Macquarie University, Sydney 2109, Australia
e-mail: pdavies@els.mq.edu.au

†Department of Astrophysics, University of New South Wales, Sydney 2052, Australia

3. Webb, J. K. *et al.* *Phys. Rev. Lett.* **87**, 091301 (2001).
4. Murphy, M. T. *et al.* *Mon. Not. R. Astron. Soc.* **327**, 1208–1222 (2001).
5. Barrow, J. D. *Phys. Rev. D* **59**, 043515 (1999).
6. Magueijo, J. *Phys. Rev. D* **62**, 103521 (2000).
7. Sandvik, H. B., Barrow, J. D. & Magueijo, J. *Phys. Rev. Lett.* **88**, 031302 (2002).
8. Bekenstein, J. D. *Phys. Rev. D* **25**, 1527–1539 (1982).
9. Barrow, J. D. & Magueijo, J. *Phys. Lett. B* **443**, 104–110 (1998).
10. Bekenstein, J. D. *Phys. Rev. D* **9**, 3292–3300 (1974).
11. Hawking, S. W. *Phys. Rev. D* **13**, 191–197 (1976).
12. Magueijo, J. *Phys. Rev. D* **63**, 043502 (2001).
13. Penrose, R. *Rev. Nuovo Cimento* **1**, 252–276 (1969).

Kinematics

Gliding flight in the paradise tree snake

Most vertebrate gliders, such as flying squirrels, use symmetrically paired 'wings' to generate lift during flight, but flying snakes (genus *Chrysopelea*) have no such appendages or other obvious morphological specializations to assist them in their aerial movements^{1–6}. Here I describe the three-dimensional kinematics of gliding by the paradise tree snake, *Chrysopelea paradisi*, which indicate that the aerial behaviour of this snake is unlike that of any other glider and that it can exert remarkable control over the direction it takes, despite an apparent lack of control surfaces.

I determined the full three-dimensional gliding trajectory of wild-caught *C. paradisi*, a southeast Asian arboreal colubrid. Snakes were videotaped and photographed jumping from a horizontal branch at the top of a 10-metre-high tower in an open field at the Singapore Zoological Gardens. Two video cameras were positioned to record in stereo, allowing the three-dimensional coordinates

of the head, midpoint and vent of the snake to be monitored throughout its trajectory.

C. paradisi prepares for take-off by hanging from a branch, with the anterior body looped into a 'J' shape. The snake jumps by accelerating up and away from the branch, straightening the body and dorsoventrally flattening it from head to vent. Its body width roughly doubles, with the ventral surface acquiring a slightly concave shape. As the snake gains speed while falling, the body pitches downwards and the head and vent are brought towards the midpoint to form an 'S' shape in the horizontal plane. The snake begins to undulate laterally, starting with the anterior body. The flight trajectory shallows (Fig. 1) as lift is generated. Throughout the trajectory, its body posture changes in a characteristic way during each undulatory cycle.

In a typical glide, the snake took off with a maximum upward acceleration of $14.4 \pm 0.8 \text{ m s}^{-2}$ (mean $\pm$ s.e.m.) and horizontal velocity of $1.7 \pm 0.1 \text{ m s}^{-1}$ ($n=7$ for both) when fully airborne. During mid-glide, the snake undulated at a frequency of $1.3 \pm 0.1 \text{ Hz}$, with a wave height (peak to trough) of $33 \pm 2\%$ snout–vent length ($n=7$ for both). The airspeed (the speed along its trajectory) and sinking speed were 8.1 ± 0.2 and $4.7 \pm 0.5 \text{ m s}^{-1}$ ($n=8$), respectively. The glide angle late in the trajectory was $31 \pm 3^\circ$ ($n=8$), although the glide angle continued to change throughout.

C. paradisi is surprisingly adept at aerial manoeuvring. In contrast to many fliers, *C. paradisi* turns without banking. Instead, turns are initiated by movement of the anterior body, and occur only during the half of the undulatory cycle when the head is moving towards the direction of the turn. In one sequence, a snake (snout–vent length, 47 cm;

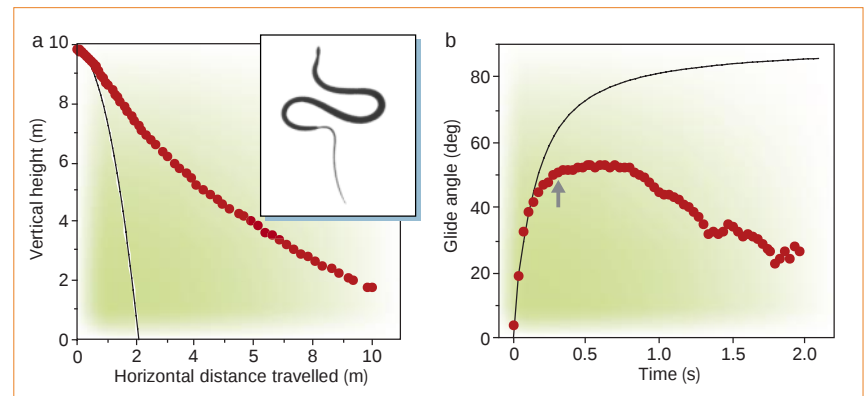


Figure 1 Representative glide trajectory of *Chrysopelea paradisi* (snout–vent length, 64 cm; mass, 27 g). **a**, Aerial trajectory, not including take-off sequence. *Chrysopelea* data points (red) are unsmoothed three-dimensional midpoint coordinates, projected in a lateral plane and sampled at 30 Hz. The wing loading, determined from the ventral silhouette of the aerial snake, is 26 N m^{-2} . The angle of attack of the anterior body ranges from 5° to 60° . Black curve, theoretical path of a projectile launched with the same initial velocity as the snake. The snake exits from the field of view of the cameras before landing. **b**, Glide angle through time for this same trajectory (red). Glide angle is the angle between the local trajectory and the horizon. The glide angle starts near zero, reflecting the snake's initial horizontal velocity, and then deviates from the theoretical projectile (black) early in the trajectory, roughly where the snake starts undulating (arrow). The glide angle levels off in the ballistic dive at about 53° , then decreases at a rate of 22° s^{-1} . Both the glide angle and the airspeed change throughout the trajectory, indicating that equilibrium was not reached in this sequence.

1. Webb, J. K. *et al.* *Phys. Rev. Lett.* **82**, 884–887 (1999).
2. Songaila, A. & Cowie, L. L. *Nature* **398**, 667–668 (1999).

mass, 11 g) avoided a tree in mid-glide by turning at an angular velocity of 0.84 rad s^{-1} .

Despite its unconventional flight behaviour, *C. paradisi*'s aerial performance is on a par with that of other gliders. Its best glide ratio (the ratio of horizontal distance gained to height lost) is 3.7, which is comparable with that of flying squirrels (*Petaurista petaurista*, 4.7)⁷, flying lizards (*Draco melanopogon*, 3.7)⁸ and flying frogs (*Rhacophorus nigropalmatus*, 2.1)⁹. *C. paradisi* is thus potentially capable of using aerial locomotion effectively to move between trees, chase aerial prey or avoid predators.

C. paradisi's aerial lateral undulation is a modified form of a more typical ophidian terrestrial locomotion, although in air the frequency is one-third lower (relative to the same snake; $n=4$) and the amplitude is higher. The timing of the start of lateral undulation in relation to the shallowing of the trajectory suggests that lateral undulation helps to generate the snake's lift. Aerial locomotion in snakes is probably more complicated than terrestrial locomotion because gliding involves lateral undulation while simultaneously maintaining a concave ventral shape; to my knowledge, this combination of movement and postural regulation is not known to occur together in any other snake and probably requires specialized neuromuscular control.

John J. Socha

Department of Organismal Biology and Anatomy,
University of Chicago, 1027 East 57th Street,
Chicago, Illinois 60637, USA

e-mail: jjsocha@midway.uchicago.edu

1. Daly, M. *Bombay Nat. Hist. J.* **12**, 589 (1899).
2. Flower, S. S. *Proc. Zool. Soc. Lond.* 16 May (1899).
3. Wall, F. J. *Bombay Nat. Hist. Soc.* **18**, 227–243 (1908).
4. Vaughn-Arbuckle, K. H. *J. Bombay Nat. Hist. Soc.* **56**, 640–642 (1959).
5. Pendlebury, H. M. *Bull. Raffles Mus.* **5**, 75 (1931).
6. Heyer, W. R. & Pongsapipatana, S. *Herpetologica* **26**, 317–319 (1970).
7. Scholey, K. D. in *Biona Report 5, Bat flight – Fledermausflug* (ed. Nachtigall, W.) 187–204 (Fischer, Stuttgart, 1986).
8. McGuire, J. *Phylogenetic Systematics, Scaling Relationships, and the Evolution of Gliding Performance in Flying Lizards (genus Draco)*. Thesis, Univ. Texas, Austin (1998).
9. Emerson, S. B. & Koehl, M. A. R. *Evolution* **44**, 1931–1946 (1990).

Competing financial interests: declared none.

COMMUNICATIONS ARISING

Nitrogen cycle

What governs nitrogen loss from forest soils?

Nitrogen is lost as dissolved organic compounds in stream waters from unpolluted South American forests, but it is lost mainly as inorganic nitrate in streams flowing from North American forests that suffer nitrogen deposition from the atmosphere¹. From this it has been inferred that the standard thinking about

how nature deals with nitrogen in soils and waters² needs to be re-evaluated and that the conventional wisdom of how nitrogen is absorbed and released by plants³ must be wrong. We disagree, however, on the grounds that there are other, more likely interpretations of the new results¹.

How nature deals with nitrogen depends greatly on temperature. Rye-grass plants supplied with equal concentrations of ammonium and nitrate take up an increasing proportion of ammonium as the temperature becomes cooler⁴. Plants are equipped with transport mechanisms for a variety of nitrogen-containing organic solutes⁵ and they can absorb small organic molecules such as amino acids in northern temperate forests with cool temperatures⁶.

This flexibility might have evolved because the microbes responsible for releasing soil organic nitrogen as ammonium, and for converting the ammonium to nitrate, become less active as the temperature falls: the conversion to nitrate is inhibited^{7,8} at 3–5 °C. This implies that the cooler the average temperature is, the more important it becomes for plants to be able to manage without nitrate and to utilize nitrogen compounds that have not been fully processed by the soil microbes.

The mean annual temperatures at the sites of the South American forest studies¹ were quite low (4–11 °C) so plants there might well have absorbed small nitrogen-containing organic molecules. But the dissolved organic nitrogen found in forest streams does not prove this: 'dissolved' was defined¹ as passing through a filter of pore size smaller than 1 µm and would therefore have included molecules up to 1,000 times larger than those taken up by plants, together with colloidal organic matter and bacteria⁹.

The 'dissolved' organic nitrogen is probably in those streams for the simple reason that it is not needed. A forest ecosystem with no input of nitrogen would evolve to recycle usable nitrogen, inorganic or organic, and to minimize its loss in streams. But very large organic molecules and colloidal organic matter are not usable by plants. The significance of the 'dissolved' organic nitrogen in those streams is not that these are the forms of nitrogen that the forest uses, but that they are the forms that it does not recycle because it cannot use them. These results do not call for a re-evaluation of our thinking about how nature deals with nitrogen in soils and waters because they are what we would expect from our current understanding of the situation.

Tom Addiscott, Phil Brookes

Rothamsted Research, Harpenden,
Hertfordshire AL5 2JQ, UK

e-mail: tom.addiscott@bbsrc.ac.uk

1. Perakis, S. S. & Hedin, L. O. *Nature* **415**, 416–419 (2002).
2. van Breemen, N. *Nature* **415**, 381–382 (2002).

3. Pearce, F. *New Scientist* **11** (26 January 2002).
4. Clarkson, D. T. & Warner, A. J. *Plant Physiol.* **64**, 557–561 (1979).
5. Williams, L. E. & Miller, A. J. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **52**, 659–688 (2001).
6. Persson, J. & Näsaholm, T. *Ecol. Lett.* **4**, 434–438 (2001).
7. Anderson, O. E. *Soil Sci. Soc. Am. Proc.* **24**, 286–289 (1960).
8. Tyler, K. B. *et al.* *Soil Sci.* **87**, 123–129 (1959).
9. Herbert, B. E. & Bertsch, P. M. in *Carbon Forms and Functions in Forest Soils* (eds Kelly, J. A. & McFee, W. W.) 63–88 (Soil Sci. Soc. Am., Madison, Wisconsin, 1995).

van Breemen replies — The predominance of organic nitrogen in stream waters and soil solutions is no proof of plant uptake of organic nitrogen, and could indeed be brought about by the uptake of only inorganic nitrogen, as Addiscott and Brookes claim and standard thinking would have it. Nor did I suggest otherwise¹. Yet I maintain that "some standard thinking about how nature deals with nitrogen in soils and waters needs to be re-evaluated".

Standard thinking is best summarized by published diagrams of the terrestrial nitrogen cycle — with one exception² that I know of, such representations in recent soil-science textbooks^{3–6} ignore two features of the nitrogen cycle that have come to light: dissolved organic nitrogen as a potentially important loss term for soil nitrogen⁷, and the apparently widespread ability of plants (including crop plants) to take up dissolved organic nitrogen^{8,9}.

Addiscott and Brookes suggest that dissolved organic nitrogen reaching stream water is rather inert. Maybe so, but it has hitherto been largely ignored and we know little about it. The free amino acids present in low concentrations in soil and stream waters probably reflect a small, dynamic pool⁸ on the way from a large pool of dissolved high-molecular-mass organic nitrogen to microorganisms, plants or ammonium. Plants might get a better share of that pool than we once thought.

Nico van Breemen

Laboratory of Soil Science and Geology,
Department of Environmental Sciences,
Wageningen University and Research Centre,
PO Box 37, 6700 AA Wageningen, The Netherlands
e-mail: nico.vanbreemen@bodeco.beng.wau.nl

1. van Breemen, N. *Nature* **415**, 381–382 (2002).
2. Fisher, F. F. & Binkley, D. *Ecology and Management of Forest Soils* 3rd edn (Wiley, New York, 2000).
3. Brady, N. C. & Weil, R. R. *The Nature and Properties of Soils* 12th edn (Prentice Hall, Upper Saddle River, New Jersey, 1999).
4. Singer, M. J. & Muns, D. N. *Soils: An Introduction* 4th edn (Prentice Hall, Upper Saddle River, New Jersey, 1999).
5. Rowell, D. L. *Soil Science, Methods and Applications* (Addison Wesley Longman, Harlow, UK, 1995; reprinted 1997).
6. Miller, R. W. & Gardiner, D. T. *Soils in our Environment* 8th edn (Prentice Hall, Upper Saddle River, New Jersey, 1998).
7. Perakis, S. S. & Hedin, L. O. *Nature* **415**, 416–419 (2002).
8. Lipson, D. & Näsaholm, T. *Oecologia* **128**, 305–316 (2001).
9. Persson, J. & Näsaholm, T. *Ecol. Lett.* **4**, 434–438 (2001).

Editorial note: See also addendum from S. S. Perakis and L. O. Hedin on page 665 of this issue.

Structural changes in the calcium pump accompanying the dissociation of calcium

Chikashi Toyoshima & Hiromi Nomura

Institute of Molecular and Cellular Biosciences, The University of Tokyo, Bunkyo-ku, Tokyo 113-0032, Japan

In skeletal muscle, calcium ions are transported (pumped) against a concentration gradient from the cytoplasm into the sarcoplasmic reticulum, an intracellular organelle. This causes muscle cells to relax after cytosolic calcium increases during excitation. The Ca^{2+} ATPase that carries out this pumping is a representative P-type ion-transporting ATPase. Here we describe the structure of this ion pump at 3.1 Å resolution in a Ca^{2+} -free (E2) state, and compare it with that determined previously for the Ca^{2+} -bound (E1 Ca^{2+}) state. The structure of the enzyme stabilized by thapsigargin, a potent inhibitor, shows large conformation differences from that in E1 Ca^{2+} . Three cytoplasmic domains gather to form a single headpiece, and six of the ten transmembrane helices exhibit large-scale rearrangements. These rearrangements ensure the release of calcium ions into the lumen of sarcoplasmic reticulum and, on the cytoplasmic side, create a pathway for entry of new calcium ions.

P-type ion transporting ATPases, which include Na^+K^+ -ATPase and gastric H^+K^+ -ATPase among others, are fundamental in establishing ion gradients by pumping ions across biological membranes (reviewed in ref. 1). Of many P-type ATPases known today, Ca^{2+} -ATPase (SERCA1a) from skeletal muscle sarcoplasmic reticulum (SR) is structurally and functionally the best-studied member^{1–3}. SR Ca^{2+} -ATPase pumps Ca^{2+} from the cytoplasm into the reticulum, thereby causing the relaxation of muscle cells. Two Ca^{2+} ions can be transported per ATP hydrolysed and two or three H^+ ions are counter-transported⁴.

Active transport of Ca^{2+} -ATPase is achieved, according to the E1–E2 model^{5,6}, by changing the affinity of Ca^{2+} -binding sites from high (E1) to low (E2)⁷. The release of Ca^{2+} ‘occluded’ in the transmembrane binding sites takes place during the transition from E1P to E2P (‘P’ indicating that the enzyme is phosphorylated; Fig. 1, inset). Autophosphorylation of an aspartyl residue in the reaction cycle is a characteristic feature of the P-type ATPases. However, the residues constituting the phosphorylation site are shared by the members in the haloacid dehalogenase superfamily⁸ and by many bacterial response regulators, despite the differences in the folding patterns (reviewed in ref. 9).

We previously determined the structure of SR Ca^{2+} -ATPase¹⁰ with two bound Ca^{2+} in the transmembrane (M) region, which consists of ten α -helices (Fig. 1; Protein Data Bank, PDB, code 1EUL). The cytoplasmic part of Ca^{2+} -ATPase consists of three domains (A, actuator or anchor; N, nucleotide; and P, phosphorylation), well separated in this Ca^{2+} -bound (E1 Ca^{2+}) form. The phosphorylation residue, Asp 351, is located on the P domain, and the adenosine moiety of ATP binds to the N domain. Modelling the structures (PDB codes 1FQU (ref. 10) and 1KJU (ref. 11)) based on low-resolution maps of the tubular crystals^{11,12}, in which the enzyme is in a state similar to E2P (ref. 13), showed large movements of the three cytoplasmic domains to form a compact headpiece. These rearrangements of the cytoplasmic domains must be associated with the changes in the transmembrane binding sites¹⁴, but the molecular mechanism was far beyond what we could imagine.

Here we describe the crystal structure of Ca^{2+} -ATPase in the absence of Ca^{2+} and in the presence of thapsigargin, a potent inhibitor that fixes the enzyme in a form analogous to E2 (ref. 15), abbreviated as E2(TG). The structure (Figs 1 and 2), deter-

mined to 3.1 Å resolution, is very different from that of E1 Ca^{2+} , yet can be compared directly, because no ATP or phosphorylation is involved in the transition between them. The movements of cytoplasmic domains are even larger than we described for the tubular crystals¹⁰. Transmembrane helices undergo drastic rearrangements that involve shifts normal to the membrane. These movements have clear mechanistic implication in the release and binding of Ca^{2+} . Knowing the second structure in the reaction cycle, we can now begin to understand how ion pumps work.

Structure determination

The crystals of SR Ca^{2+} -ATPase were grown by dialysis in the presence of thapsigargin and exogenous lipid. Thapsigargin was added before the addition of EGTA for removing Ca^{2+} . Two types of crystals of different symmetry ($P2_1$ and $P4_1$) appeared but only those belonging to the space group $P4_1$ diffracted to 3.1 Å resolution at the BL44XU beam line of SPring-8. The structure was solved by molecular replacement and refined to an R_{free} of 26.8%. Inclusion of Ca^{2+} (1 mM) in the dialysis buffer did not affect the crystal quality or bring differences in the final model. The structure of Ca^{2+} -ATPase in E2(TG) form, in comparison with E1 Ca^{2+} , is shown in Figs 1 and 2. We describe the structural changes as those accompanying the transition from E1 Ca^{2+} to E2(TG), mainly because we prepared the crystals by removing Ca^{2+} .

Movements of the cytoplasmic domains

To realize the close association of the three cytoplasmic domains widely separated in E1 Ca^{2+} , the N domain inclines nearly 90° with respect to the membrane (Fig. 2a) and the A domain rotates by about 110° horizontally (Fig. 2b; Supplementary Information Animation 3). As a result, the top part of the N domain moves more than 50 Å. Also, the trypsin digestion site on the A domain (T2 in Fig. 1), which is well exposed in E1 Ca^{2+} , is partially blocked by the P domain¹³ (Supplementary Information Animation 3). The cytoplasmic domains move as a whole in an M10-to-M1 direction (~23 Å for the P domain).

Two components contribute to this large change in inclination of the N domain: the first is the movement of the P domain (see, for example, the change in inclination of the P5 helix in Fig. 2a), and the other is that of the N domain itself relative to the P domain. An

analysis by Dyndom¹⁶ indicates that the P domain inclines by about 30° with respect to the membrane, and the N domain by about 50° relative to the P domain. This 30° inclination of the P domain is directly related to tilting of the transmembrane helices, in particular the M5 helix, and is likely to be a central event in the active transport (see below).

Despite the large movements between the two states, the structure of the P and N domains remain virtually the same (r.m.s. deviations 0.63 and 0.75 Å, respectively; Fig. 3 and Supplementary Information Animation 3), except for a few 'hinge' residues and those at the A–P interface. The hinge region (DPPR starting from Asp 601 and TNQMS starting from Thr 358) is well hydrogen bonded between the two strands; no hydrogen bonds favouring E2(TG) can be identified, although it is generally difficult to assign hydrogen bonds at 3.1 Å resolution. This suggests that stabilization of the closed configuration takes place entirely at the A–N and A–P interfaces, where we can identify a few hydrogen bonds (Fig. 1). The residues at the A–P interface (Ile 179–Thr 181 in A, and Asp 703–Asn 706 in P) contain signature sequences of the P-type ATPases (reviewed by ref. 1).

Rearrangement of transmembrane helices

The dissociation (or binding) of Ca²⁺ accompanies dramatic rearrangements of six (M1–M6) out of ten transmembrane helices (Figs 2 and 4a). The rearrangements are not limited to those constituting the Ca²⁺-binding sites (M4–M6, M8) and appear quite complicated: M1 and M2 move upwards (+z direction, towards the cytoplasm) whereas M3 and M4 shift downwards,

both up to about 5 Å, close to one turn of an α-helix (5.4 Å; Figs 1 and 2). M1 shows a large lateral movement within the membrane (Fig. 2b), and is unique in this respect. M3 and M5 are strongly curved in E2(TG) but in opposite directions (Figs 1 and 2). Our first goal is, therefore, to understand how these complicated movements are organized and linked with those of the cytoplasmic domains. We can start with the many places where a movement of one helix could cause the movements of others (a 'domino' effect). For example, in Fig. 4a, inclination of M4 must impinge on M2 at the top, causing M2 itself to tilt. It also helps to note that the movements of helices normal to the membrane are in most cases produced by the changes in their inclinations (for example, M2). However, it seems most critical to understand the links between the transmembrane helices and the P domain.

M5 runs through the centre of the enzyme from the luminal surface to the end of the P domain, where it is integrated as a part of the Rossmann fold (Fig. 3). Hence, the top part of M5 moves together with the P domain as a single entity. In fact, if the two structures are superimposed with the P domain, the top part of M5 also superimposes virtually completely (Fig. 3). Here, short anti-parallel β-strands (0 and 7) of the P domain appear to 'clamp' the M4 and M5 helices (Fig. 3). The middle part of M5 is linked to the loop (L67) connecting M6 and M7 (through Arg 751; Fig. 1), perhaps restricting the bowing of M5; this loop is also hydrogen bonded to the P domain (refer to Fig. 8 of ref. 10) and is important in phosphorylation¹⁷. On the opposite side of the P domain to M4, M3 is located. The top part of M3 is connected to the P1 helix at the bottom of the P domain through a critical hydrogen bond¹⁸

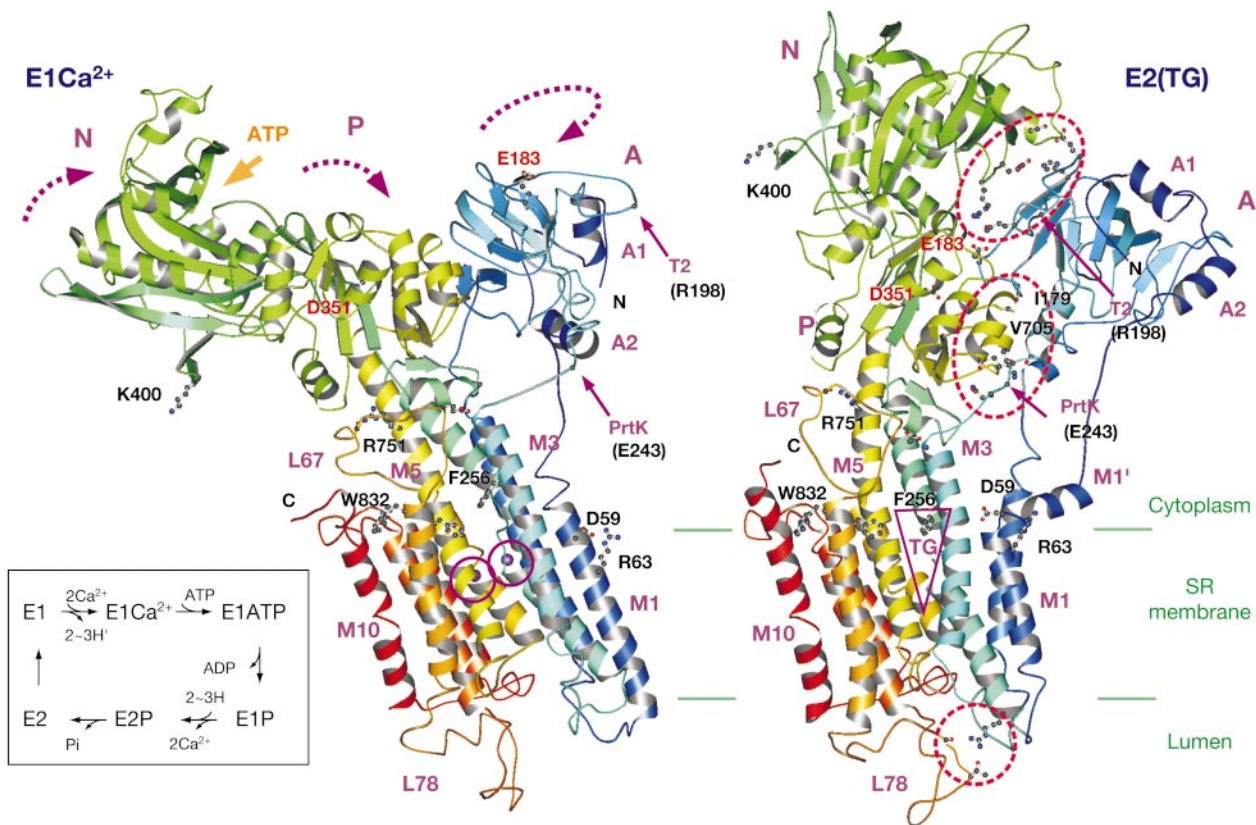


Figure 1 Ribbon representation of SR Ca²⁺-ATPase in the Ca²⁺-bound form (E1Ca²⁺) and that (E2(TG)) in the absence of Ca²⁺ but in the presence of thapsigargin (TG). Inset, a simplified reaction scheme (showing only the forward direction). Colours change gradually from the amino terminus (blue) to the carboxy terminus (red). Two purple spheres (circled) in E1Ca²⁺ represent bound Ca²⁺. Red circles in E2(TG) indicate extra hydrogen bonds in E2(TG). Large arrows in E1Ca²⁺ indicate the direction of movement of the cytoplasmic

domains during the change from E1Ca²⁺ to E2(TG). PrtK, proteinase-K digestion site (around Glu 243; ref. 27); T2, trypsin digestion site at Arg 198 (ref. 41); ATP, binding pocket for the adenosine moiety of ATP. Principal residues are marked: E183 (A domain), F256 (thapsigargin-binding site), D351 (P domain, phosphorylation site), K400 (N domain, phospholamban-binding site⁴²) and R751 (linking M5 and the loop (L67) connecting M6 and M7). Prepared with Molscript⁴³.

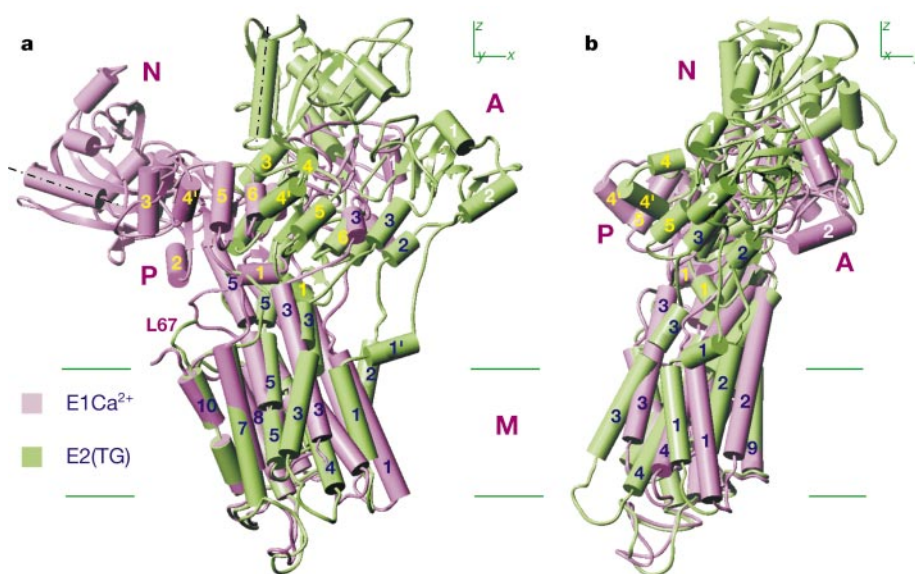


Figure 2 Superimposition of the Ca^{2+} -bound form (E1Ca^{2+} , violet) and the thapsigargin-bound form (E2(TG) , light green) of Ca^{2+} -ATPase fitted with the transmembrane domain. α -Helices are represented by cylinders and β -strands by arrows. Both are viewed along the membrane plane, but from nearly orthogonal directions (specified in the top right

corners). M5 is represented with three cylinders, and M3 with two cylinders, although they are continuous helices. Dash-dotted lines in **a** show the orientations of a helix in the N domain in the two states. Also see Supplementary Information Animations 1 and 2.

involving Glu 340 (Fig. 3).

Thus, M3–M5 helices are directly linked to the P domain by hydrogen bonds. M6 is also connected, although less directly, to the P domain through L67 (Fig. 4a) and M5 (by hydrogen bonds between Asn 756 and the carbonyl of Gly 808 and Asn 810). If the P domain inclines, for instance, owing to the bending of M5, all of these helices (M3–M6) will incline and generate movements that have components normal to the membrane. Their amounts depend on the distances from the pivoting point (Fig. 4b), which is located around Gly 770 (on M5), a critical residue¹⁹, at the middle of the membrane (double circle on M5 in Fig. 4a). There, M5 is tightly packed against M7 with a cluster of four (one on M5, three on M7) glycines, and the lower part below Gly 770 hardly moves (Fig. 4a;

Supplementary Information Animation 4). The shift is therefore small for M6 and large for M3 and M4 (Figs 2 and 4a); whole M3 and M4 helices move downwards during dissociation of Ca^{2+} , whereas M6 undergoes more-local changes (Supplementary Information Fig. 1).

Because tilting of the P domain will cause different movements in different parts of the protein, interfaces between them will change. This might result in steric clashes, which can be avoided only by adjusted movements. For instance, if the M3 helix moved as a single entity with the P domain, the luminal end would collide with M5 (Fig. 2a). The large $-y$ component of the movement of M3 (Fig. 2b) must be to avoid this (or is guided by the steric constraints). The movement of M4 is also explained in this way. To be able to change the interface between transmembrane helices, hydrogen bonds between them should be avoided, which is certainly the case. The ‘joints’ also need to be flexible. In fact, M3–M5 helices do not actually move with the P domain as a single entity. Only the top part of M5 does so; M4 differs in orientation even at the top (Fig. 3). Presumably this is because the β -strands clamping M5 and M4 (0 and 7; Fig. 3) are short and located at an end of the β -sheet in the Rossmann fold, allowing more freedom. The link between the P domain and M3 involves the long side chain of Glu 340 (Fig. 3). Thus, the P domain, with flexible joints, functions as a coordinator of the transmembrane helices that gather by mostly van der Waals interactions.

It is now easy to see that the P domain will conduct the movements of M1 and M2, although they are not directly connected. M1 and M2 are characterized by their upward movements with Ca^{2+} dissociation. The movement of M1 is complex: the top part of M1 is largely bent at Asp 59 (Fig. 1) and forms an amphipathic helix having hydrophobic residues (Phe, Leu and Ile) on one side and charged residues (Glu 58, Asp 59) on the other (M1' in Fig. 2). This bending may have occurred as a result of steric collision with M3, because M3 inclines and move downwards whereas M1 moves upwards. M2 is an inclined helix in E1Ca^{2+} but nearly upright in E2(TG) (Fig. 4a). This change in inclination, apparently caused by M4 (Fig. 4a), results in a large ($\sim 5 \text{ \AA}$) upward movement near the cytoplasmic surface but hardly any at the luminal end. In contrast, M1 moves upwards without changing its inclination, but also moves

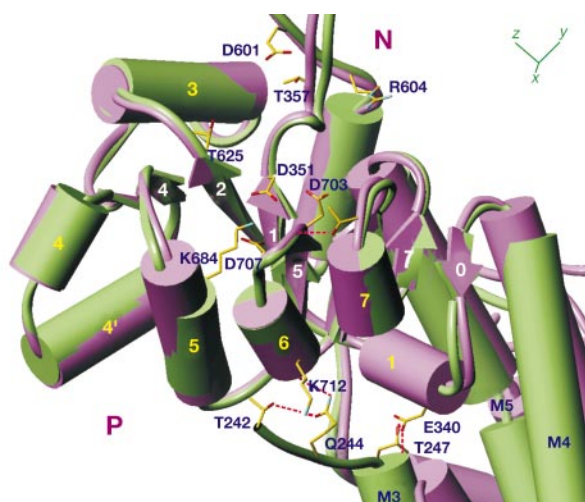


Figure 3 Interface between the transmembrane helices (M3–M5) and the P domain of Ca^{2+} -ATPase. Superimposition of the Ca^{2+} -bound (E1Ca^{2+} , violet) and thapsigargin-bound (E2(TG) , light green) forms fitted with the P domain. The residues (in atom colour) represent those in E2(TG) . Links between the P1 and M3 helices involve hydrogen bonds between E340 and NH of L249 (not seen) as well as OH of T247 near the top of M3. Also see Supplementary Information Animation 3.

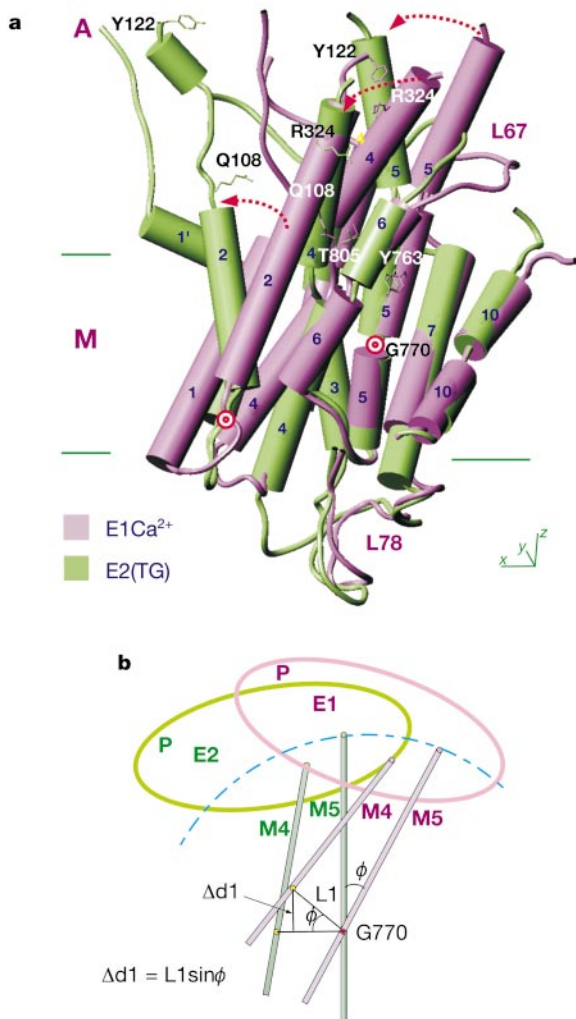


Figure 4 Rearrangement of transmembrane helices viewed from the rear (**a**), and a diagram illustrating the shift of M4 normal to the membrane by the tilting of M5 (**b**). The models for E1Ca²⁺ (violet) and E2(TG) (light green) are superimposed. The M5 helix lies along the plane of the paper. M8 and M9 are removed in **a**. Double circles show pivot positions for M2 and M5. Arrows indicate the directions of movements during the change from E1Ca²⁺ to E2(TG). In **b**, M4 and M5 are linked in the P domain (ovals) and move as a rigid body. Tilting of M5 around the pivoting point (Gly 770) generates a vertical shift of $\Delta d1$ for M4 without a large horizontal shift at the level of the pivoting point. For $L1 = 10 \text{ \AA}$ and $\phi = 30^\circ$, $\Delta d1$ becomes $5 \text{ \AA}$. Also see Supplementary Information Animation 4.

horizontally (pushed by M2; Fig. 2b). The top part of M2 seems to be positioned by Tyr 122 with a hydrogen bond to Arg 324 (M4) in E1Ca²⁺ (Fig. 4a). This will break with the inclination of M4. In E2(TG), M2 is partly unwound but the top part remains, packed against the P domain (Fig. 4a). Thus, although M1 and M2 are connected to the A domain and expected to be important in its rotation, their movements are likely to be governed by the P domain.

Changes around the Ca²⁺-binding sites

These movements of transmembrane helices have clear meanings in ion binding and dissociation. Figure 5 shows a more detailed view around the Ca²⁺-binding sites. The two sites (I and II) have different coordination geometry¹⁰ and biochemical properties²⁰. Site I, presumably the binding site for the first Ca²⁺ ion^{20,21}, entirely consists of side-chain oxygen atoms (Fig. 5a) of residues on three helices (M5, M6 and M8). Site II, with an arrangement of oxygen atoms reminiscent of the EF-hand motif²², is formed nearly 'on' the

M4 helix; three carbonyl oxygen atoms from residues on M4 contribute, and Glu 309 caps the bound Ca²⁺ from above (Fig. 5a). To realize efficient coordination of Ca²⁺, both M4 and M6 helices have unwound parts (Supplementary Information Fig. 1).

In site I, we see marked changes with the residues on M6. Three crucial residues, Asn 796, Thr 799 and Asp 800, rotate nearly 90° clockwise (orange arrows in Fig. 5c) accompanying the dissociation of Ca²⁺. As a result, Thr 799 orients away from the molecule's centre, and is replaced by Asp 800 (Fig. 5c). Asn 768 moves away towards site II. The backbone of Glu 771 and Glu 908 hardly move. In short, the change in inclination of M5 above Gly 770 (Fig. 4a) decreases the number of oxygen atoms that can contribute to site I. In site II, Glu 309 now points away from the binding site (Fig. 5b), M4 is shifted downwards, and Asp 800 moves away (Fig. 5c). The carbonyl groups involved in Ca²⁺ binding may make hydrogen bonds with Asn 768 and Asn 796, both of which have moved closer to M4.

This configuration of M6 in E2(TG) seems to be stabilized by a hydrogen bond between Ser 940 (M9) and the main-chain carbonyl of Thr 799 (Fig. 5b), which was hydrogen bonded to the hydroxyl of Tyr 763 in E1Ca²⁺. The carbonyl oxygen of Asp 800, binding water in E1Ca²⁺, seems to make a hydrogen bond with the amide of Ala 804 (Supplementary Information Fig. 1). Ser 767 may work as the hydrogen bond donor for Glu 771 and Asp 800. At this resolution, however, hydrogen-bonding networks remain ambiguous, particularly because no water molecules can be identified.

These changes clearly explain the decrease of affinity for Ca²⁺, although it is unclear why such complicated movements are required. Homology modelling of the cation-binding sites of Na⁺K⁺-ATPase (H. Ogawa and C.T., unpublished data) provides a clue. With the arrangements of residues shown here, it was straightforward to make two high-affinity K⁺-binding sites, as long as Asn 796 is replaced by Asp, after the Na⁺K⁺-ATPase sequence: the other coordinating residues are common to both ATPases. The key feature in the model is that the Asp (Asn 796) is coordinated to both K⁺, similar to Asp 800 in coordination of two Ca²⁺. Because Asn 796 is located one turn below where Asp 800 was, M4 must move downwards to provide carbonyl groups for coordination of K⁺. It is well established that K⁺ affects many kinetic parameters of Ca²⁺-ATPase, although in millimolar range (for example, ref. 23). We therefore suggest that large complicated movements are needed for counter-transport. It would be interesting to know what ions are counter-transported by the pump protein ancestral¹ to both Ca²⁺- and Na⁺K⁺-ATPases.

Thapsigargin-binding site

It was necessary to include thapsigargin in the specimen to keep the enzyme from denaturing in the absence of Ca²⁺. Thapsigargin has attracted considerable interest because of its high affinity (subnanomolar dissociation constant¹⁵) and specificity. Thapsigargin is a hydrophobic molecule that is thought to bind to the M3 helix around Phe 256 (refs 24, 25). Our study unambiguously identifies its binding site: the cavity surrounded by the M3, M5 and M7 helices near the cytoplasmic surface of the membrane (Fig. 6). Bulky hydrophobic residues on these three helices form a complementary surface to thapsigargin, which, with a potential hydrogen bond with Ile 829 (Fig. 6), effectively reduces the movements of transmembrane helices. In E1Ca²⁺ this cavity is narrower (Fig. 2a) and the surface is not complementary because of the shift of M3.

Why thapsigargin can save solubilized Ca²⁺-ATPase from denaturation is worth considering. The instability suggests that thermal movements of transmembrane helices become deleteriously large when left without the support of lipids or the cohesion provided by bound Ca²⁺. In fact, phospholipids and millimolar Ca²⁺ are necessary for long-term stability of this enzyme²⁶. High Ca²⁺ concentrations will work as a kind of inhibitor that limits the movement of transmembrane helices essential for Ca²⁺ transport.

Requirement of millimolar Ca^{2+} has also been demonstrated by protection to the cytoplasmic domains against proteinase K attack²⁷. In Na^+K^+ - and H^+K^+ -ATPases, transmembrane helices come out of the membrane after proteolysis and mild heat treatment^{28,29}. This happens only in the absence of K^+ . These results suggest that transmembrane helices undergo large-scale movements when binding cations are absent, and that the movements are linked with those of the cytoplasmic domains.

Access pathway to the binding sites

In the classical E1–E2 models, E2 is described as having low-affinity Ca^{2+} -binding sites exposed to the lumen, and Ca^{2+} binding from the cytoplasmic side requires prior conversion to an E1 form⁷. Therefore we do not know whether we should expect to see, in the E2(TG) structure, an ion pathway to the transmembrane binding sites. Yet a water-accessible channel clearly exists that has an opening ($8 \times 10 \text{ \AA}$) lined by negatively charged residues and leads to the carboxyl group of Glu 309 (Supplementary Information Fig. 2). It is therefore tempting to speculate that the first event in the Ca^{2+} binding is an interaction with Glu 309, followed by a conformation change of the side chain to deliver the Ca^{2+} to the binding cavity. Indeed, there is no room for Ca^{2+} to pass around Glu 309 to reach the binding sites. This channel disappears in E1Ca^{2+} owing to the movements of M1 and M3.

On the luminal side, the loop connecting M3 and M4 comes closer to the L78 loop (Fig. 2a), sealing the access pathway at the very surface. In E1Ca^{2+} , the pathway is much deeper, coming close to the Ca^{2+} -binding sites. Thus, in this aspect, the structure deviates from the classical E1–E2 models that assume lumenally opened binding

sites⁷. However, in the normal reaction cycle, the luminal gate is presumably opened during the E1P-to-E2P transition, and the counter ions (H^+ in this case) bind then (Fig. 1, inset). Hence, in E2, because the counter ions have already bound, it will be more efficient for a pump to close the luminal gate and prepare for the entry of new Ca^{2+} from the cytoplasmic side. These features are exactly what we see in the structure presented here.

Conclusion

As described, Ca^{2+} -ATPase seems to undergo, in the absence of Ca^{2+} , large-scale thermal movements involving both transmembrane and cytoplasmic domains, and the P domain is the coordinator of the movements. If so, we expect the SR vesicles with loaded Ca^{2+} to show significant leakage, and thapsigargin to stop it. This has indeed been demonstrated (G. Inesi, personal communication). Certainly, the closed configuration of the cytoplasmic domains will limit such movements and therefore the leakage. Another role of the closed configuration is to stop the reaction cycle, which is regulated essentially by Ca^{2+} alone. ATP can bind to the enzyme even when Ca^{2+} is absent but, without Ca^{2+} , the reaction cycle cannot proceed. In the closed configuration, the γ -phosphate of ATP comes close to, but cannot reach, the phosphorylation residue Asp 351. It requires deeper inclination of the N domain, or the rotation of the A domain to release the N domain. This in turn requires the binding of Ca^{2+} , which will fix the transmembrane helices in different positions and will prevent the tilting of the P domain to form a closed configuration. In this regard, it is noteworthy that the A-domain rotation takes place during the E1P-to-E2P transition¹⁴, and that the proteolytic cut of the link

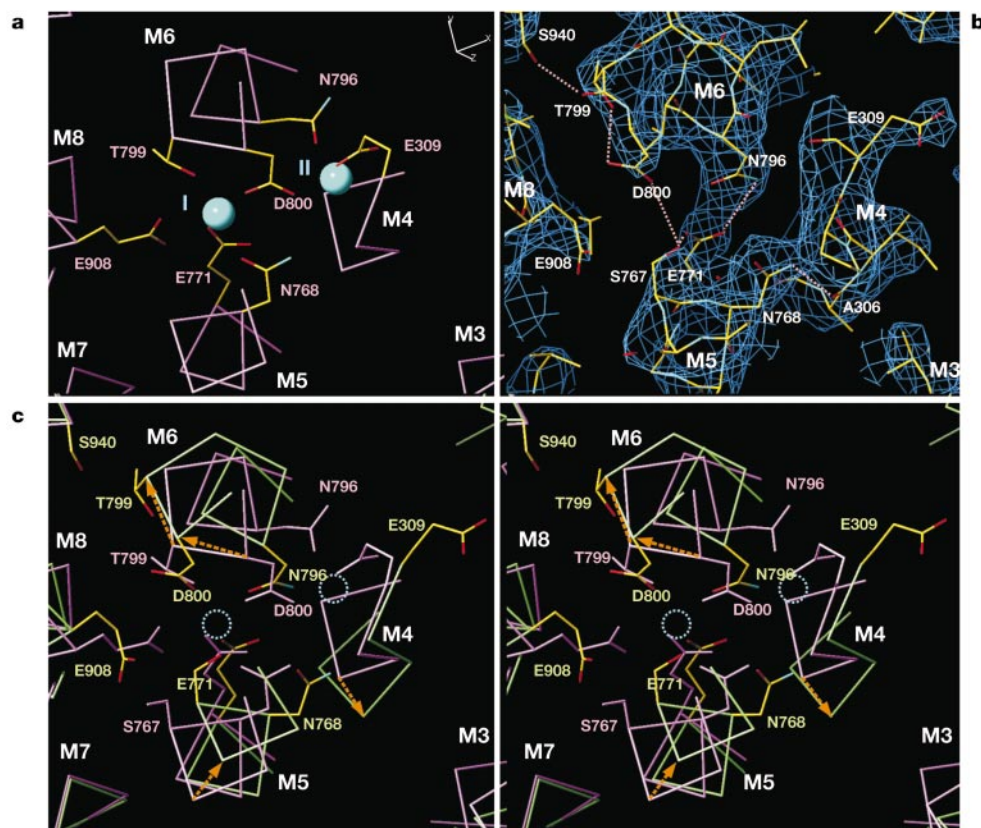


Figure 5 Conformation changes around the Ca^{2+} -binding sites. **a**, Co trace and the side chains of the coordinating residues in E1Ca^{2+} . **b**, Composite omit map³⁷ (at 1.2σ) and the model for the corresponding area in E2(TG). **c**, Stereo view of the composite of the models for E1Ca^{2+} (violet) and E2(TG) (atom colour). The viewing direction is approximately down

the M5 helix in E1Ca^{2+} . Two bound Ca^{2+} appear as cyan spheres (**a**) or circles (**c**). Dashed lines in **b** show potential hydrogen bonds. Orange arrows in **c** show the movements of the corresponding residues during the change from E1Ca^{2+} to E2(TG).

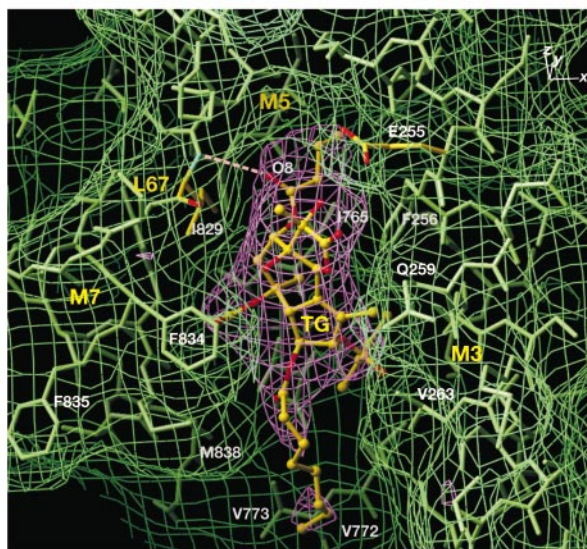


Figure 6 Thapsigargin (TG) binding site in Ca^{2+} -ATPase. TG is in atom colour, and Ca^{2+} -ATPase in green (except for 1829 and E255). The nets represent the omit-annealed $F_o - F_c$ map for TG (at 3σ , violet) and the water-accessible surface of Ca^{2+} -ATPase (green). The dashed pale-pink line shows a potential hydrogen bond between O8 of TG and NH of 1829. The orientation of TG is consistent with the differences in affinity observed with natural variants and derivatives⁴⁴.

between the A domain and M3 results in complete inhibition of the ATPase activity^{30,31}, suggesting that the rotation of the A domain is another main event in the active transport by this pump. □

Methods

Crystallization

Affinity-purified enzyme^{32,33} (20 μM) in detergent octaethyleneglycol mono-*n*-dodecylether (C_{12}E_8 , 2 mg ml^{-1}) and in 1 mM Ca^{2+} was mixed with thapsigargin (30 μM) and 3 mM EGTA, then dialysed against a buffer consisting of 2.75 M glycerol, 4% polyethylene glycol (PEG 400), 3 mM MgCl_2 , 2.5 mM $\text{Na}_2\text{S}_2\text{O}_8$, 2 $\mu\text{g ml}^{-1}$ butylhydroxytoluene, 0.2 mM dithiothreitol, 0.1 mM EGTA, 20 mM MES buffer at pH 6.1, for about 1 month. Crystals were grown to $200 \times 200 \times 50 \mu\text{m}$ and flash-frozen in cold nitrogen gas.

Data collection

Diffraction data were collected at BL44XU, SPring-8, using a DIP2040 imaging plate detector ($4,000 \times 4,000$ pixels; MAC Science) at $\lambda = 0.90 \text{ \AA}$, and processed with Denzo and Scalepack³⁴. Data from two of the best crystals were merged ($R_{\text{merge}} = 9.2\%$ with a redundancy of 7.1 and $I/\sigma = 21.2$) for $1/15.0$ to $1/3.1 \text{ \AA}^{-1}$ (43.7%, 4.1, 3.0, respectively, for the highest resolution bin, $3.18\text{--}3.10 \text{ \AA}$). The crystals belonged to space group $P4_1$ (unit cell dimensions of $a = b = 71.7 \text{ \AA}$ and $c = 590.3 \text{ \AA}$) and contained two protein molecules in the asymmetric unit.

Modelling and refinement

The modelling was started from a model¹⁰ for the cytoplasmic part of the enzyme in the tubular crystals¹². Initial refinement³⁵ was done with the N and P domains, then the A domain was included. The rest of the structure was gradually constructed by many iterations of map drawing, solvent flattening with Solomon³⁶ and model building using Turbo-Frodo (BioGraphics). The model, consisting of 7,717 atoms (7,673 in the protein, 44 in thapsigargin), was refined with CNS³⁷ using 50,822 reflections (at 95.7% completeness) to an R_{free} of 0.268 ($R_{\text{cryst}} = 0.237$); r.m.s. bond length and angle were $0.009 \text{ \AA}$ and 1.3° , respectively. Strict non-crystallographic symmetry constraints were applied throughout. The geometry of the model was examined with Procheck³⁶ and no residues (except for Gly) had dihedral angles in the disallowed region. The secondary structures were assigned with DSSP³⁸. Crystal structures of trilobolide³⁹ and an epoxide of thapsigargin⁴⁰ were used as the initial model of thapsigargin.

Received 4 April; accepted 27 June 2002; doi:10.1038/nature00944.

- Möller, J. V., Juul, B. & le Maire, M. Structural organization, ion transport, and energy transduction of P-type ATPases. *Biochim. Biophys. Acta* **1286**, 1–51 (1996).
- MacLennan, D. H., Rice, W. J. & Green, N. M. The mechanism of Ca^{2+} transport by sarco(endo)plasmic reticulum Ca^{2+} -ATPases. *J. Biol. Chem.* **272**, 28815–28818 (1997).
- Lee, A. G. & East, J. M. What the structure of a calcium pump tells us about its mechanism. *Biochem. J.* **356**, 665–683 (2001).
- Yu, X., Carroll, S., Rigaud, J. L. & Inesi, G. H^+ countertransport and electrogenicity of the

- sarcoplasmic reticulum Ca^{2+} pump in reconstituted proteoliposomes. *Biophys. J.* **64**, 1232–1242 (1993).
- Post, R. L., Hegyvary, C. & Kume, S. Activation by adenosine triphosphate in the phosphorylation kinetics of sodium and potassium ion transport adenosine triphosphatase. *J. Biol. Chem.* **247**, 6530–6540 (1972).
- Albers, R. W. Biochemical aspects of active transport. *Annu. Rev. Biochem.* **36**, 727–756 (1967).
- de Meis, L. & Vianna, A. L. Energy interconversion by the Ca^{2+} -dependent ATPase of the sarcoplasmic reticulum. *Annu. Rev. Biochem.* **48**, 275–292 (1979).
- Aravind, L., Galperin, M. Y. & Koonin, E. V. The catalytic domain of the P-type ATPase has the haloacid dehalogenase fold. *Trends Biochem. Sci.* **23**, 127–129 (1998).
- Johnson, L. N. & Lewis, R. J. Structural basis for control by phosphorylation. *Chem. Rev.* **101**, 2209–2242 (2001).
- Toyoshima, C., Nakasako, M., Nomura, H. & Ogawa, H. Crystal structure of the calcium pump of sarcoplasmic reticulum at $2.6 \text{ \AA}$ resolution. *Nature* **405**, 647–655 (2000).
- Xu, C., Rice, W. J., He, W. & Stokes, D. L. A structural model for the catalytic cycle of Ca^{2+} -ATPase. *J. Mol. Biol.* **316**, 201–211 (2002).
- Zhang, P., Toyoshima, C., Yonekura, K., Green, N. M. & Stokes, D. L. Structure of the calcium pump from sarcoplasmic reticulum at $8 \text{ \AA}$ resolution. *Nature* **392**, 835–839 (1998).
- Danko, S. *et al.* ADP-insensitive phosphoenzyme intermediate of sarcoplasmic reticulum Ca^{2+} -ATPase has a compact conformation resistant to proteinase K, V8 protease and trypsin. *FEBS Lett.* **489**, 277–282 (2001).
- Danko, S., Yamasaki, K., Daiho, T., Suzuki, H. & Toyoshima, C. Organization of cytoplasmic domains of sarcoplasmic reticulum Ca^{2+} -ATPase in E₁P and E₁ATP states: a limited proteolysis study. *FEBS Lett.* **505**, 129–135 (2001).
- Sagara, Y. & Inesi, G. Inhibition of the sarcoplasmic reticulum Ca^{2+} transport ATPase by thapsigargin at subnanomolar concentrations. *J. Biol. Chem.* **266**, 13503–13506 (1991).
- Hayward, S. Structural principles governing domain motions in proteins. *Proteins* **36**, 425–435 (1999).
- Zhang, Z., Lewis, D., Sumbilla, C., Inesi, G. & Toyoshima, C. The role of the M6-M7 loop (L67) in stabilization of the phosphorylation and Ca^{2+} binding domains of the sarcoplasmic reticulum Ca^{2+} -ATPase (SERCA). *J. Biol. Chem.* **276**, 15232–15239 (2001).
- Zhang, Z. *et al.* Mutational analysis of the peptide segment linking phosphorylation and Ca^{2+} -binding domains in the sarcoplasmic reticulum Ca^{2+} -ATPase. *J. Biol. Chem.* **270**, 16283–16290 (1995).
- Andersen, J. P., Vilsen, B. & MacLennan, D. H. Functional consequences of alterations to Gly310, Gly770, and Gly801 located in the transmembrane domain of the Ca^{2+} -ATPase of sarcoplasmic reticulum. *J. Biol. Chem.* **267**, 2767–2774 (1992).
- Zhang, Z. *et al.* Detailed characterization of the cooperative mechanism of Ca^{2+} binding and catalytic activation in the Ca^{2+} transport (SERCA) ATPase. *Biochemistry* **39**, 8758–8767 (2000).
- Andersen, J. P. & Vilsen, B. Amino acids Asn796 and Thr799 of the Ca^{2+} -ATPase of sarcoplasmic reticulum bind Ca^{2+} at different sites. *J. Biol. Chem.* **269**, 15931–15936 (1994).
- Glusker, J. P. Structural aspects of metal liganding to functional groups in proteins. *Adv. Protein Chem.* **42**, 1–76 (1991).
- Medda, P., Fassold, E. & Hasselbach, W. The effect of monovalent and divalent cations on the ATP-dependent Ca^{2+} -binding and phosphorylation during the reaction cycle of the sarcoplasmic reticulum Ca^{2+} -transport ATPase. *Eur. J. Biochem.* **165**, 251–259 (1987).
- Yu, M. *et al.* Specific substitutions at amino acid 256 of the sarcoplasmic/endoplasmic reticulum Ca^{2+} transport ATPase mediate resistance to thapsigargin in thapsigargin-resistant hamster cells. *J. Biol. Chem.* **273**, 3542–3546 (1998).
- Hua, S. & Inesi, G. Synthesis of a radioactive azido derivative of thapsigargin and photolabeling of the sarcoplasmic reticulum ATPase. *Biochemistry* **36**, 11865–11872 (1997).
- Pikula, S., Mullner, N., Dux, L. & Martonosi, A. Stabilization and crystallization of Ca^{2+} -ATPase in detergent-solubilized sarcoplasmic reticulum. *J. Biol. Chem.* **263**, 5277–5286 (1988).
- Juul, B. *et al.* Do transmembrane segments in proteolyzed sarcoplasmic reticulum Ca^{2+} -ATPase retain their functional Ca^{2+} binding properties after removal of cytoplasmic fragments by proteinase K? *J. Biol. Chem.* **270**, 20123–20134 (1995).
- Lutsenko, S., Anderko, R. & Kaplan, J. H. Membrane disposition of the M5-M6 hairpin of Na^+ , K^+ -ATPase α subunit is ligand dependent. *Proc. Natl Acad. Sci. USA* **92**, 7936–7940 (1995).
- Gatto, C., Lutsenko, S., Shin, J. M., Sachs, G. & Kaplan, J. H. Stabilization of the H,K-ATPase M5M6 membrane hairpin by K^+ ions. Mechanistic significance for P2-type ATPases. *J. Biol. Chem.* **274**, 13737–13740 (1999).
- Juul, B. & Möller, J. V. In *Na/K-ATPase and Related ATPases* (eds Taniguchi, K. & Kaya, S.) 233–236 (Elsevier, Amsterdam, 2000).
- Jørgensen, P. L. & Collins, J. H. Tryptic and chymotryptic cleavage sites in sequence of α -subunit of ($\text{Na}^+ + \text{K}^+$)-ATPase from outer medulla of mammalian kidney. *Biochim. Biophys. Acta* **860**, 570–576 (1986).
- Coll, R. J. & Murphy, A. J. Purification of the CaATPase of sarcoplasmic reticulum by affinity chromatography. *J. Biol. Chem.* **259**, 14249–14254 (1984).
- Stokes, D. L. & Green, N. M. Three-dimensional crystals of CaATPase from sarcoplasmic reticulum. Symmetry and molecular packing. *Biophys. J.* **57**, 1–14 (1990).
- Otwiński, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–325 (1997).
- Brünger, A. T. Extension of molecular replacement: a new search strategy based on Patterson correlation refinement. *Acta Crystallogr. A* **46**, 46–57 (1990).
- Collaborative Computational Project No. 4 The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
- Brünger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
- Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
- Kutschabsky, L., Kretschmer, R.-G. & Ripperger, H. The crystal and molecular structure of the sesquiterpenoid silerin (trilobolide). *Crystal Res. Technol.* **21**, 627–633 (1986).
- Christensen, S. B., Larsen, I. K., Rasmussen, U. & Christopherson, C. Thapsigargin and thapsigargin, two histamine liberating sesquiterpene lactones from *Thapsia garganica*. X-ray analysis of the 7,11-epoxide of thapsigargin. *J. Org. Chem.* **47**, 649–652 (1982).

41. MacLennan, D. H., Brandl, C. J., Korczak, B. & Green, N. M. Amino-acid sequence of a Ca^{2+} + Mg^{2+} -dependent ATPase from rabbit muscle sarcoplasmic reticulum, deduced from its complementary DNA sequence. *Nature* **316**, 696–700 (1985).
42. James, P., Inui, M., Tada, M., Chiesi, M. & Carafoli, E. Nature and site of phospholamban regulation of the Ca^{2+} pump of sarcoplasmic reticulum. *Nature* **342**, 90–92 (1989).
43. Kraulis, P. J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
44. Christensen, S. B., Andersen, A. & Smitt, U. W. Sesquiterpenoids from *Thapsia* species and medicinal chemistry of the thapsigargin. *Fort. Chem. Org. Nat.* **71**, 129–167 (1997).

Supplementary Information is available on *Nature's* website (<http://www.nature.com/nature>).

Acknowledgements

We thank H. Ogawa for help in data gathering, R. Yoshida for computations, and M. Nakasako for modelling. We also thank G. Inesi, P. Champeil, D. B. McIntosh and

H. Suzuki for communicating unpublished results to us and for their help in improving the manuscript. Thanks are also due to E. Yamashita and all the staff at BL44XU of SPring-8. This work was supported in part by Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology, the Japan New Energy and Industry Technology Development Organization, and the Human Frontier Science Program.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.T. (e-mail: ct@iam.u-tokyo.ac.jp). The atomic coordinates are deposited in the PDB (accession code 1IWO).

Coherent properties of a two-level system based on a quantum-dot photodiode

A. Zrenner^{*†}, E. Beham^{*}, S. Stufli^{*†}, F. Findeis^{*}, M. Bichler^{*} & G. Abstreiter^{*}

^{*} Walter Schottky Institut, Technische Universität München, Am Coulombwall, D-85748 Garching, Germany

Present-day information technology is based mainly on incoherent processes in conventional semiconductor devices¹. To realize concepts for future quantum information technologies, which are based on coherent phenomena, a new type of 'hardware' is required². Semiconductor quantum dots are promising candidates for the basic device units for quantum information processing. One approach is to exploit optical excitations (excitons) in quantum dots. It has already been demonstrated that coherent manipulation between two excitonic energy levels—via so-called Rabi oscillations—can be achieved in single quantum dots by applying electromagnetic fields^{3–7}. Here we make use of this effect by placing an InGaAs quantum dot in a photodiode, which essentially connects it to an electric circuit. We demonstrate that coherent optical excitations in the quantum-dot two-level system can be converted into deterministic photocurrents. For optical excitation with so-called π -pulses, which completely invert the two-level system, the current is given by $I = fe$, where f is the repetition frequency of the experiment and e is the elementary charge. We find that this device can function as an optically triggered single-electron turnstile.

Semiconductor quantum dots (QDs), often referred to as artificial atoms, are suitable entities with which to implement arrays of quantum bits (qubits) for quantum information processing. One possible approach to this is the use of excitonic excitations in the ground state of a QD as the basis of a two-level system (Fig. 1a). Rabi oscillations have been successfully demonstrated in the exciton population of single QDs^{3–5}, showing that the abstract model of a two-level system is in fact applicable to the excitonic ground state of QDs. Low-temperature dephasing times for excitons in self-assembled QDs have been shown to exceed several hundred picoseconds, allowing sufficiently high numbers of coherent manipulations with picosecond pulses^{8,9}.

Here we report the coherent optoelectronic properties of a single QD in a photodiode, which is essentially a single two-level system with electric contacts. Coherent optical excitations of this system are expected to result in Rabi oscillations, which can be directly and quantitatively transferred into photocurrents¹⁰. For our experiments we fabricated single-QD photodiodes starting from a layer of self-assembled In_{0.5}Ga_{0.5}As QDs grown by molecular beam epitaxy. Although based on a conventional diode structure, here a GaAs n-i-Schottky structure, the only optically active part is a single self-assembled In_{0.5}Ga_{0.5}As QD contained in the intrinsic layer of the diode (Fig. 1b). A semitransparent Schottky contact is provided by a 5-nm-thick titanium layer. The optical selection of a single QD is done by shadow masks with apertures from 100 to 500 nm, which are prepared by electron beam lithography from a 80-nm-thick aluminium layer (see ref. 11 for further details). A schematic diagram of such a diode is shown in Fig. 1c. The QD represents a single two-level system, which is designed to interact with a coherent optical pump pulse (impinging through the shadow mask) by excitonic transitions, and with an electric circuit by tunnelling. It is this connection to the electric circuit which offers

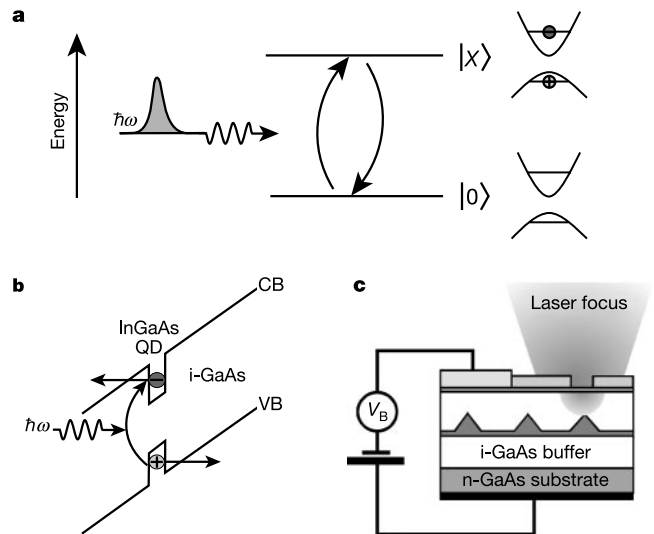


Figure 1 The two-level system in a quantum dot, and the single-quantum-dot photodiode. **a**, Schematic view of an excitonic two-level system in the ground state of a semiconductor QD. The state $|0\rangle$ corresponds to an empty QD, $|X\rangle$ to a one-exciton occupancy. Transitions between both states can be induced by resonant optical driving fields, where the energy $\hbar\omega$ of the optical excitation corresponds to the energetic level spacing between the states $|0\rangle$ and $|X\rangle$. Coherent optical π -pulse excitation in particular then results in an inversion of the two-level system with respect to its initial state: $|0\rangle \leftrightarrow |X\rangle$. **b**, Intrinsic region of a biased single-QD photodiode: Coherent optical π -pulse excitation applied to the initial state $|0\rangle$ results in the final state $|X\rangle$, which corresponds to the coherent generation of a single electron-hole pair. The photoionization of this state by tunnelling results in the separation of a single electron-hole pair and hence in a deterministic photocurrent. **c**, Schematic view of a single-QD photodiode on the basis of a GaAs n-i-Schottky diode. Optical access to a single QD is provided by a shadow mask. CB, conduction band; VB, valence band.

those optoelectronic features described above.

We use linear photocurrent spectroscopy (Ti:sapphire laser) with continuous wave (c.w.) excitation to characterize our single-QD photodiode^{11,12}. With a fixed laser energy E_L close to the QD exciton resonance E_X , we tune the bias voltage V_B and, as a consequence of the quantum confined Stark effect (QCSE), $E_X(V_B)$ is shifted through E_L (see Fig. 2 for $E_L = 1.31$ eV). By repeating this procedure for different E_L , we obtain a magnitude of the QCSE of about 2.4 meV V^{-1} (Fig. 2 inset). The excitonic photocurrent resonance exhibits a typical width of about 30 mV, which corresponds to a spectral linewidth of about $72 \mu\text{eV}$. Owing to the nearly linear Stark shift, the V_B axis can be directly transferred to an energy axis. On the basis of this known (approximately linear) QCSE, a single-QD photodiode can be operated as detector with resonant, electric field tunable sensitivity. Within a limited energy range it can be applied as an 'all in one' mesoscopic optical spectrum analyser.

We now consider the coherent properties of a single-QD photodiode, and address the hierarchy of relevant and characteristic timescales in this system. The lifetime of a coherent polarization is given by the decoherence or dephasing time τ_{dephase} . Only within this characteristic time is it possible to observe coherent interactions or to perform coherent manipulations in the system. In the field of semiconductor QDs, experiments on self-assembled InGaAs QDs have shown low-temperature dephasing times in excess of 500 ps (refs 8, 9). The spontaneous radiative recombination lifetime of the exciton τ_{rad} is of the order of 1 ns in typical III/V semiconductor QDs with direct bandgap. It is the longest timescale relevant for excitonic two-level systems. In biased single-QD photodiodes, the tunnelling time τ_{tunnel} is also important. Depending on the electric field, τ_{tunnel} is tunable from infinity to less than 3 ps (ref. 11). For

[†] Present address: Universität Paderborn, Warburger Strasse 100, D-33098 Paderborn, Germany.

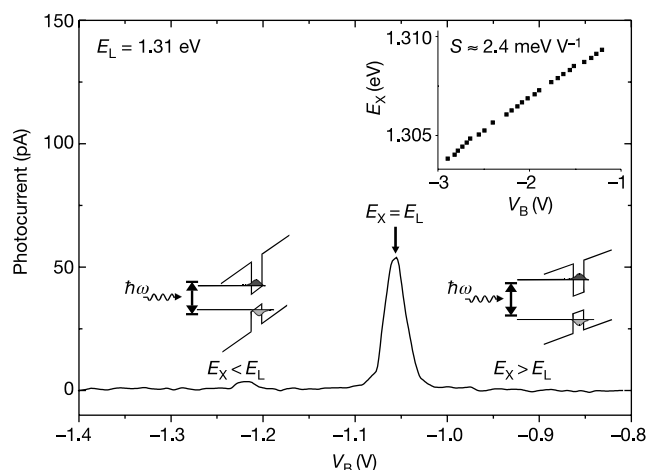


Figure 2 Photocurrent spectrum of a single QD in the region of the excitonic ground state energy E_X . In this c.w. experiment, the laser energy E_L is fixed, whereas E_X is tuned by V_B via the quantum confined Stark effect, QCSE. Within the QCSE tuning range (inset), the single-QD photodiode can be operated as a V_B tunable spectrometer. On the left side of the resonance ($E_X < E_L$), the QD appears red-shifted with respect to the fixed laser energy $E_L = \hbar\omega$, on the right side ($E_X > E_L$), it appears blue-shifted.

typical bias conditions in our present device ($V_B = -1.04$ V), we infer $\tau_{\text{tunnel}} \approx 10$ ps from the photocurrent linewidth. According to simple rate equations, the excitonic excitation decays mostly by tunnelling for $\tau_{\text{tunnel}} < \tau_{\text{rad}}$, and hence contributes to the photocurrent. Tunnelling on a timescale $\tau_{\text{tunnel}} < \tau_{\text{dephase}}$, on the other hand, leads to enhanced dephasing, and limits the available time range for coherent interactions to τ_{tunnel} . For future applications, τ_{dephase} and τ_{tunnel} could be chosen independently by introducing a time-dependent bias voltage V_B (ref. 10). In such an arrangement, coherent manipulations of the two-level system could be performed at zero electric field under the condition of weak dephasing, whereas subsequent read-out would be performed at high electric field.

To reach conditions for coherent excitation for the present case of time-independent V_B , we apply 1-ps pulses from a mode-locked Ti:sapphire laser. Within the laser line width of about 0.7 meV, no other QD state contributes to the photocurrent. With a biexciton binding energy of 3 meV, the two-photon biexciton resonance¹³ is sufficiently out of range, 1.5 meV below E_X , which corresponds to a 0.6-V separation on the V_B axis. We varied the pulse intensity over a range of up to two orders of magnitude. For the following analysis, we follow the photocurrent signal on the main resonance at about -1.04 V. In Fig. 3 we show the on-resonance photocurrent as a function of the excitation amplitude (A_{exc}), which is the square root of the peak pulse intensity, the relevant parameter for coherent phenomena. Here the relative unit of 1 corresponds to a time-averaged c.w. excitation power of about $1 \times 10^3 \text{ W cm}^{-2}$ (within a 1- μm laser focus) incident on the shadow mask of the diode, for a pulse width of about 1 ps and a pulse repetition frequency of 82 MHz. In the present work we have scaled the excitation amplitude such that $A_{\text{exc}} = 1$ corresponds to the condition of an optical π -pulse.

With increasing A_{exc} , we first observe an increasing photocurrent signal from the QD. At $A_{\text{exc}} \approx 1$ the current reaches a maximum; a further increase of A_{exc} then leads to a reduction of the photocurrent. This behaviour is in contrast to the well-known characteristics of a conventional, incoherent photodiode, but is expected for the coherent population of a two-level system with increasing pumping power^{6,7}. Over the whole range of A_{exc} we observe more than one period of a damped Rabi oscillation in the photocurrent, which reflects directly and quantitatively the resulting occupancy in

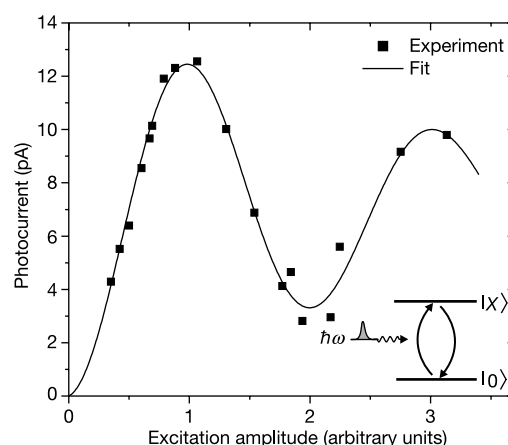


Figure 3 Rabi oscillations of the photocurrent at resonance for increasing excitation amplitude A_{exc} . Coherent π -pulse excitation corresponds to $A_{\text{exc}} = 1$. The experimentally observed peak photocurrent is 12.6 pA, only slightly below the predicted limit of $I = fe = 13.1$ pA ($f = 82$ MHz). For π -pulse excitation, the single-QD photodiode therefore performs as optically triggered single-electron turnstile device. The fitted curve assumes an exponential damping of the Rabi oscillations with increasing A_{exc} , and is included as a guide to the eye.

the two-level system. The observed Rabi oscillations appear damped at high A_{exc} , as previously observed in related work³⁻⁵. The exact nature of the underlying dephasing mechanism(s) remains an open question. The range around the first maximum of the Rabi oscillations, which corresponds to an excitation with a π -pulse ($A_{\text{exc}} = 1$), appears effectively undamped. Applied to a two-level system with an initial occupancy of 0, a single π -pulse projects the system to occupancy 1 (complete inversion). The photoionization of this excitation by tunnelling leads then to the separation of (in the ideal case) exactly one electron-hole pair, and hence to the net transport of one elementary charge between the contacts of the photodiode. In this sense a single-QD photodiode, excited with a π -pulse in the coherent regime, is therefore a deterministic current source, which delivers one elementary charge e to an outer circuit per laser pulse. With the pulse repetition frequency f of our laser (82 MHz), we expect therefore a time-integrated net current $I = fe$ of 13.1 pA. In our present experiment, we obtain a peak photocurrent in the first maximum of about 12.6 pA. This quantitative measure of the occupancy results in about 96% of the maximum possible photocurrent. The demonstration of such a high efficiency in probing the occupancy of a specific quantum-mechanical state became possible by applying the above-described concept of electric readout.

A coherently driven single-QD photodiode is an optically triggered single-electron source, which offers, just like the related single-electron turnstile device¹⁴, single electrons (or holes) on demand. Both devices have the potential to deliver frequency-controlled currents according to $I = fe$. With circularly polarized excitation, a single-QD photodiode has the potential to provide deterministic streams of spin-polarized charges. More subtle points concerning the absolute accuracy of this device, like the influence of incomplete tunnelling or possible avalanche multiplication, will certainly be the subject of future work. By means of single-QD photodiodes, it becomes possible to transfer optical excitations of single quantum systems into deterministic electric currents. This optoelectronic functionality provides a link between coherent optical excitations and electric currents on the single-electron level. We believe that this will allow the electric readout of excitonic quantum gates, and could have a substantial impact on quantum information technology. □

Received 10 April; accepted 17 June 2002; doi:10.1038/nature00912.

1. Alferov, Z. I. Nobel Lecture: The double heterostructure concept and its applications in physics, electronics, and technology. *Rev. Mod. Phys.* **73**, 767–782 (2001).
2. Bouwmeester, D., Ekert, A. & Zeilinger, A. (eds) *The Physics of Quantum Information* (Springer, Berlin, 2000).
3. Stievater, T. H. *et al.* Rabi oscillations of excitons in single quantum dots. *Phys. Rev. Lett.* **87**, 133603–1–133603–4 (2001).
4. Kamada, H. *et al.* Exciton Rabi oscillation in a single quantum dot. *Phys. Rev. Lett.* **87**, 247401–1–247401–4 (2001).
5. Htoon, H. *et al.* Interplay of Rabi oscillations and quantum interference in semiconductor quantum dots. *Phys. Rev. Lett.* **88**, 087401–1–087401–4 (2002).
6. Rabi, I. I. Space quantization in a gyrating magnetic field. *Phys. Rev.* **51**, 652–654 (1937).
7. Allen, L. & Eberly, J. H. *Optical Resonance and Two-Level Atoms* (Wiley, New York, 1975).
8. Borri, P. *et al.* Ultralong dephasing time in InGaAs quantum dots. *Phys. Rev. Lett.* **87**, 157401–1–157401–4 (2001).
9. Bayer, M. & Forchel, A. Temperature dependence of the exciton homogeneous linewidth in $\text{In}_{0.60}\text{Ga}_{0.40}\text{As}$ self-assembled quantum dots. *Phys. Rev. B* **65**, 041308–1–041308–4 (2002).
10. Zrenner, A. German patent DE 100 06 909 C1 (4 April 2001).
11. Findeis, F. *et al.* Photocurrent and photoluminescence of a single self-assembled quantum dot in electric fields. *Appl. Phys. Lett.* **78**, 2958–2960 (2001).
12. Beham, E. *et al.* Nonlinear ground-state absorption observed in a single quantum dot. *Appl. Phys. Lett.* **79**, 2808–2810 (2001).
13. Brunner, K. *et al.* Sharp-line photoluminescence and two-photon absorption of zero-dimensional biexcitons in a GaAs/AlGaAs structure. *Phys. Rev. Lett.* **73**, 1138–1141 (1994).
14. Pothier, H. *et al.* Single-electron pump based on charging effects. *Europhys. Lett.* **17**, 249–254 (1992).

Acknowledgements

This work was supported by the BMBF and by the Deutsche Forschungsgemeinschaft.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.Z.
(e-mail: zrenner@physik.upb.de).

Evidence for spin–charge separation in quasi-one-dimensional organic conductors

T. Lorenz*, M. Hofmann*, M. Grüninger*, A. Freimuth*†, G. S. Uhrig‡, M. Dumm§ & M. Dressel§

* II. Physikalisches Institut, Universität zu Köln, 50937 Köln, Germany

‡ Institut für Theoretische Physik, Universität zu Köln, 50937 Köln, Germany

§ I. Physikalisches Institut, Universität Stuttgart, 70550 Stuttgart, Germany

Interacting conduction electrons are usually described within Fermi-liquid theory¹, which states that, in spite of strong interactions, the low-energy excitations are electron-like quasiparticles with charge and spin. In recent years there has been tremendous interest in conducting systems that are not Fermi liquids, motivated by the observation of highly anomalous metallic states in various materials, most notably the copper oxide superconductors^{2,3}. Non-Fermi-liquid behaviour is generic to one-dimensional interacting electron systems, which are predicted to be Luttinger liquids^{4,5}. One of their key properties is spin–charge separation: instead of quasiparticles, collective excitations of charge (with no spin) and spin (with no charge) are formed, which move independently and at different velocities. However, experimental confirmation of spin–charge separation remains a challenge. Here we report experiments probing the charge and heat current in quasi-one-dimensional conductors—the organic Bechgaard salts^{6–10}. It was found that the charge and

spin excitations have distinctly different thermal conductivities, which gives strong evidence for spin–charge separation. The spin excitations have a much larger thermal conductivity than the charge excitations, which indicates that the coupling of the charge excitations to the lattice is important.

Quasi-one-dimensional conductors have been realized in carbon nanotubes¹¹, in nanoscopic quantum wires based on GaAs–AlGaAs heterostructures¹², in conducting polymers¹³ and in crystalline solids^{6–10}. The one-dimensionality of these systems results either from a geometrical restriction as in a quantum wire, or from the particular crystal and electronic structure of a three-dimensional material. This is the case in the Bechgaard salts $(\text{TM})_2\text{X}$, which consist of stacks of the organic molecules tetramethylselenafulvalene (where TM is TMTSF) or tetramethyltetrathiofulvalene (where TM is TMTTF) with different anions X^- such as PF_6^- and ClO_4^- . The electronic properties are strongly anisotropic, because the overlap of the relevant orbitals is much larger in the stacking direction than perpendicular to this direction. The materials therefore provide excellent model systems for (weakly coupled) chains of interacting electrons. Owing to the full charge transfer of one electron from the chains to each of the anions the band is quarter-filled (or half-filled if a small structural dimerization along the chain direction is taken into account). Because of the large Coulomb interaction between the chain-electrons, the materials must be considered to be strongly correlated Mott–Hubbard systems. The electronic phase diagram (Fig. 1) is extraordinarily rich. Various ordered states occur at low temperatures, where the weak inter-chain coupling causes a higher-dimensional behaviour. Here, we focus on the one-dimensional Luttinger-liquid behaviour expected at elevated temperatures.

To detect signatures of spin–charge separation we studied the electrical (σ) and thermal (k) conductivities of three different materials, the spin-Peierls system $(\text{TMTTF})_2\text{PF}_6$, the spin-density wave system $(\text{TMTSF})_2\text{PF}_6$, and the superconductor $(\text{TMTSF})_2\text{ClO}_4$, which we abbreviate in the following as TM-SP, TM-SDW, and TM-SC, respectively. The experiments are based on a simple

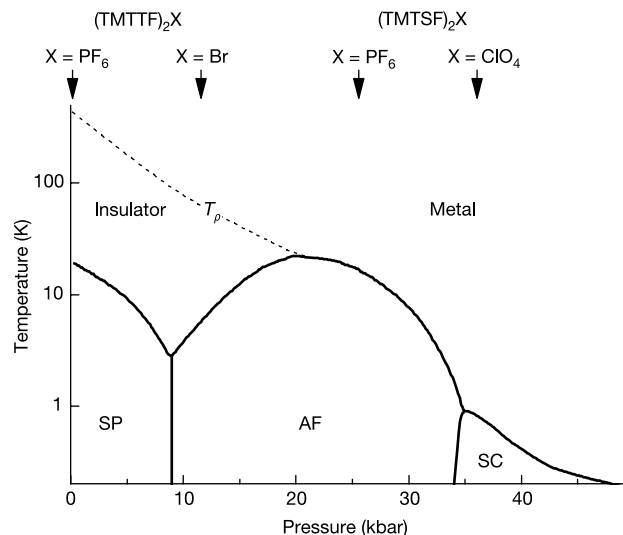


Figure 1 Phase diagram for the $(\text{TM})_2\text{X}$ compounds. Pressure or anion (X) substitution yields various phases. The arrows mark the ambient-pressure offset for each compound. The sulphur series (TMTTF) is conducting at high temperatures and insulating below the localization temperature T_p . Further cooling results in either spin-Peierls (SP) or antiferromagnetic spin density wave (AF) order. SP order is non-magnetic and results from the dimerization of the lattice and of the spins. The selenide-based salts (TMTSF) are metallic at high temperatures. At low temperatures various ordered phases are found (AF and superconducting (SC)). After ref. 10.

† Present address: Department of Physics and Astronomy, University of British Columbia, Vancouver, V6T 1Z1, Canada.

idea: whereas the electrical current probes the charge response only, the heat current is sensitive to charge and spin excitations. A comparison may therefore enable distinction between these excitations. The temperature dependence of the electrical conductivity along the chain direction (Fig. 2c inset) confirms the phase diagram of Fig. 1: TM-SP is insulating below room-temperature; TM-SDW is a metal above the spin-density wave transition at T_{SDW} and insulating for $T < T_{\text{SDW}}$; TM-SC is metallic. In contrast to the diverse behaviour of σ , we identify a common behaviour of the thermal conductivity k_a along the chain direction (Fig. 2), which is completely different from that found in conventional solids. The temperature dependence of k_a is characterized by a low-temperature maximum around 20 to 30 K and by a steep increase above 100 K. This increase is the most remarkable feature of the data. Its slope becomes smaller at high temperatures, in particular in TM-SP (Fig. 2a), which indicates a tendency towards saturation or a maximum of k_a . The thermal conductivity k_b perpendicular to the chains was measured on the same sample of TM-SP (Fig. 2a). We find $k_a(T) \approx k_b(T)$ at low temperatures, but k_b does not increase strongly above 100 K. An additional feature of the data (not shown) is a strong suppression of k_a for $T < 100$ K by magnetic fields $H||a$ and $H||b$ for all three materials ($\Delta k_a/k_a \approx 3\% T^{-1}$ at the maximum of k_a for $H \leq 15$ T). Previous studies of k_a in TM-SC in this low-temperature regime also indicate an influence of the magnetic field^{14,15}.

In an insulator such as TM-SP, heat conduction is usually due to acoustic phonons (that is, quantized sound waves). Their thermal conductivity $k_{\text{ph}}^{\text{ac}}$ has a maximum at low temperatures $T \ll \Theta_D$, the Debye temperature, and decreases strongly at higher temperatures. Here Θ_D is the typical energy of acoustic phonons. Θ_D is small, of the order of 60 K, in the Bechgaard salts¹⁶. We show in Fig. 3 a calculation of $k_{\text{ph}}^{\text{ac}}$ of TM-SP based on the standard Debye model¹⁷. The low-temperature maxima of both k_a and k_b are described well by $k_{\text{ph}}^{\text{ac}}$, but obviously not their behaviour at high temperatures. In a molecular crystal such as the Bechgaard salts, the large number of atoms in the unit cell (about 60) may give rise to an additional

contribution $k_{\text{ph}}^{\text{opt}}$ to the thermal conductivity from optical phonons. Usually $k_{\text{ph}}^{\text{opt}} \ll k_{\text{ph}}^{\text{ac}}$. This is confirmed by an estimate¹⁸ of $k_{\text{ph}}^{\text{opt}}$ of the $(\text{TM})_2\text{X}$ -salts which yields $k_{\text{ph}}^{\text{opt}} \approx 0.1\text{--}0.2 \text{ W K}^{-1} \text{ m}^{-1}$ at 300 K (Fig. 3). The weak increase of k_b at higher temperatures may be attributed to this contribution, but $k_{\text{ph}}^{\text{opt}}$ cannot account for the strong increase of k_a . Moreover, the anisotropy of the phononic heat current (in particular that of optical phonons) is not strongly temperature dependent, in contrast to our findings (Fig. 2a inset). We therefore conclude that the strong increase of k_a at high temperatures must have a non-phononic origin.

In TM-SP the chain electrons are localized, forming an insulating antiferromagnetic spin chain. The only possible heat transport mechanism next to phonons is by collective magnetic excitations. Significant heat transport by magnetic excitations at elevated temperatures, dominating the phononic heat current, was recently observed in other quasi-one-dimensional insulating spin systems^{19,20}. Common to these systems are large values of the magnetic exchange interaction J_{ex} between neighbouring spins. Notably, $J_{\text{ex}} \approx 500$ K is also large in TM-SP²¹. Therefore, a large magnetic contribution to the heat current must also be present in TM-SP and it is straightforward to attribute the high-temperature increase of k_a to a magnetic contribution k_m to the thermal conductivity. The large value of $J_{\text{ex}} \gg \Theta_D$ explains why k_m is large at high temperatures, where it dominates even the total thermal conductivity. Moreover, k_m is present only along the chain direction, which explains the anisotropy of the heat current above 100 K. An estimate of k_m is obtained using the kinetic equation, $k_m = c_m v_m l_m$. We take $v_m = (\pi^2 k_B / h) J_{\text{ex}} a \approx 7 \times 10^4 \text{ m s}^{-1}$ for the velocity of spin excitations ($k_B = 1.38 \times 10^{-23} \text{ J K}^{-1}$ is the Boltzmann constant, $h = 6.63 \times 10^{-34} \text{ J s}$ is the Planck constant, and $a \approx 0.7$ nm denotes the lattice constant of the chains) and for simplicity a temperature-independent mean free path l_m . Then k_m has the same temperature dependence as the specific heat c_m . In a one-dimensional antiferromagnet c_m has a maximum at about $J_{\text{ex}}/2$ (ref. 22). For TM-SP, $J_{\text{ex}}/2 \approx 250$ K consistent with the high-temperature behaviour of k_a . Setting $l_m = 2.8$ nm ($\approx 4a$) yields the correct magnitude of k_m

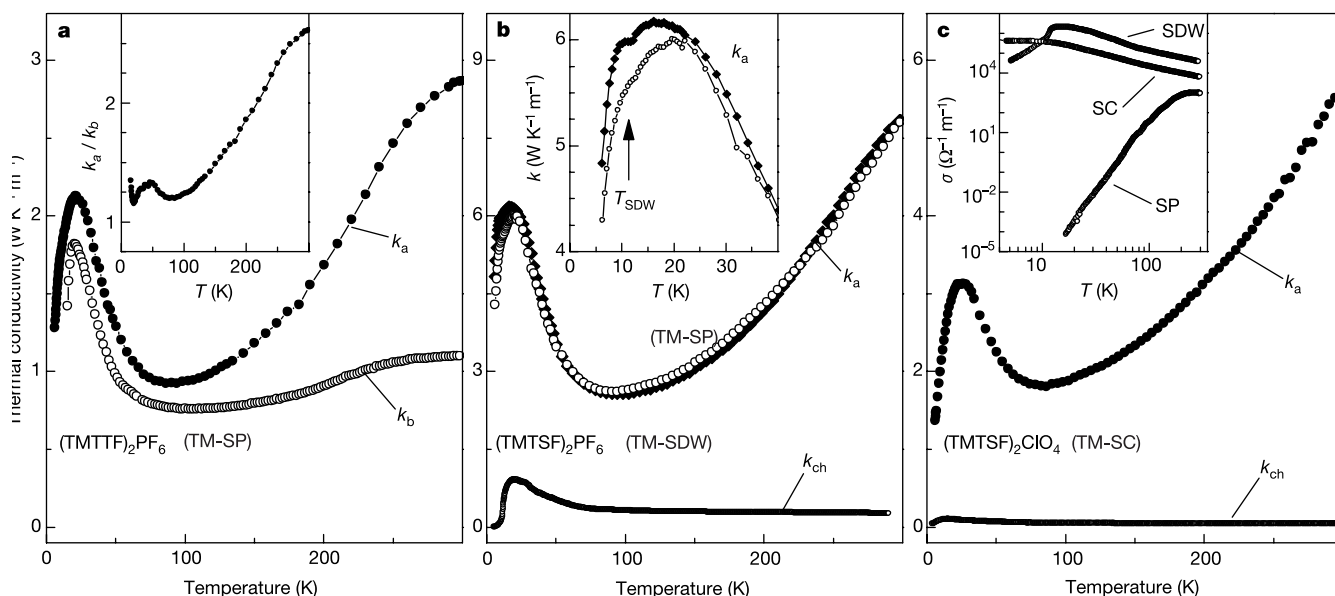


Figure 2 Thermal (k) and electrical (σ) conductivity of $(\text{TM})_2\text{X}$. **a–c**, Thermal conductivity $k_a(T)$ of $(\text{TM})_2\text{X}$ along the chains. **a**, TM-SP; **b**, TM-SDW; **c**, TM-SC. The excellent reproducibility of the temperature dependence of k_a is exemplified for TM-SDW (**b**), where we show $k_a(T)$ for two different samples (S1: filled diamonds; S2: open circles) scaled onto each other by a factor of 0.3, that is, $k_a(\text{S2}) \approx 0.3 k_a(\text{S1})$. $k_b(T)$ is the thermal conductivity perpendicular to the chains (approximately along the b-direction, which is at

80° with respect to the a-direction of the triclinic crystals). k_{ch} is the thermal conductivity of the charge carriers deduced from the electrical conductivity using the Wiedemann–Franz law (see text). Inset to **a**, anisotropy ratio k_a/k_b versus T for TM-SP. Inset to **b**, $0.3 k_a$ (sample S1) and k_a (sample S2) close to the spin-density wave transition at T_{SDW} . Inset to **c**, electrical conductivities $\sigma(T)$ along the chains of all three compounds.

(Fig. 3). These arguments give strong evidence that the high-temperature increase of k_a in TM-SP arises from a magnetic contribution to the heat current along the chains.

We turn now to the metallic compounds. Almost the same thermal conductivity as in the insulating spin-Peierls material (Fig. 2a) occurs in the conductors (Fig. 2b, c). This suggests that also in the metallic systems the heat transport is dominated by phonons and magnetic excitations. What about heat transport by charge excitations, which usually dominates the thermal conductivity of a metal? We estimate the charge contribution k_{ch} to the thermal conductivity from the electrical conductivity using the Wiedemann–Franz law¹⁷, $k_{ch} = L_0 T \sigma$, where $L_0 \approx 2.45 \times 10^{-8} \text{ W } \Omega \text{ K}^{-2}$ is the Lorenz number. This law is expected to hold for elastic scattering for excitations carrying charge and heat. Our estimate yields $k_{ch} \ll k_a$ at all temperatures (Fig. 2b, c). This is strongly corroborated by the very weak anomaly of k_a at the metal–insulator transition at T_{SDW} in TM-SDW. It is straightforward to attribute the small jump-like decrease of k_a at T_{SDW} (Fig. 2b inset) to the vanishing of the charge contribution in the insulating state below T_{SDW} . This yields an estimate of the charge contribution to the thermal conductivity of the order $0.5 \text{ W K}^{-1} \text{ m}^{-1} \ll k_a$, in agreement with what is found from the Wiedemann–Franz law. Thus, the contribution of the charge excitations to the total heat current is clearly unimportant and cannot explain the thermal conductivity of the metallic systems.

The above arguments suggest that even in the metals, the magnetic excitations carry most of the heat at high temperatures. The similarity of k_m in the metallic systems and in the spin-Peierls insulator implies that, in spite of the presence of mobile charge carriers, the chains behave like a one-dimensional spin system with respect to their magnetic properties. Moreover, the magnetic heat current is not hampered by the presence of the charge excitations, and the spin and charge contributions to k_a have a different temperature dependence and magnitude. This is clear evidence for the separation of spin and charge excitations.

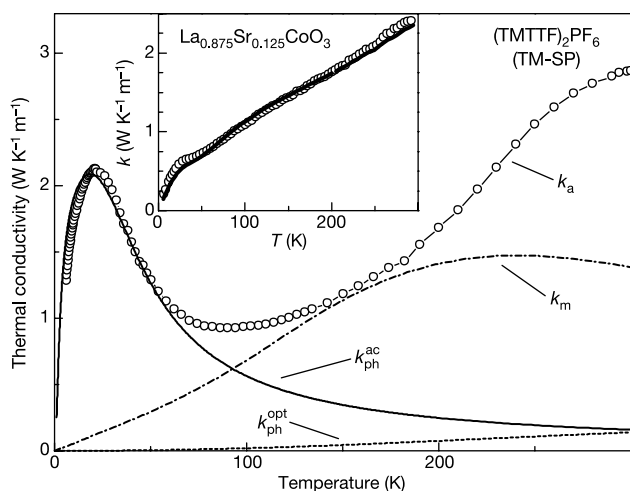


Figure 3 Thermal conductivity $k_a(T)$ parallel to the chains of the insulating spin-Peierls compound $(\text{TMTTF})_2 \text{PF}_6$. k_{ph}^{ac} is the thermal conductivity of the acoustic phonons in a Debye model¹⁷. k_{ph}^{opt} is an estimate of the thermal conductivity of 58 equidistant dispersionless optical phonon branches (each representing one longitudinal and two transverse modes) with energies between 50 and 1,500 K (ref. 18). A rough estimate of the magnetic contribution to the thermal conductivity k_m is based on $k_m = c_m v_m / m$ with c_m and v_m for a spin-1/2 Heisenberg chain (see text). To extract $I_m(T)$ a more elaborate theoretical approach is needed. Inset, thermal conductivity of $\text{La}_{0.875}\text{Sr}_{0.125}\text{CoO}_3$ measured on a large sample (solid line) and on a sample comparable in size, shape and magnitude of the thermal conductivity to the crystals of the $(\text{TM})_2\text{X}$ -salts (circles). Radiation losses are smaller than 1% (see Methods).

Our results give additional information about the decoupled spin and charge excitations. The finding $k_{ch} \ll k_m$ implies that the charge current is suppressed more strongly than the spin current, at least for $T > 100 \text{ K}$. The most likely reason for this suppression is that the charge excitations couple much more strongly to elastic deformations of the lattice, that is, to phonons, than the spin excitations²⁶. A strongly suppressed charge transport is consistent with the behaviour of the optical conductivity, in which the coherent charge excitations (the Drude peak) carry only 1% of the total spectral weight²³. This indicates a small effective velocity of the charge carriers (about 1/100 of the bare Fermi velocity). In other words, the effective band width W^* for coherent charge excitations is reduced by a factor of about 100 from the bare value⁶ $W \approx 10,000 \text{ K}$, so that the energy scale of coherent charge excitations is much lower than that of the spin excitations. In particular, for $T > W^* \approx 100 \text{ K}$ the entropy of coherent charge excitations is almost entirely released, that is, almost all charge excitations are incoherent, which implies a strongly suppressed conductivity in the charge channel.

Our results give evidence for spin–charge separation and thus for non-Fermi-liquid behaviour and the concept of a Luttinger-liquid state in one-dimensional conductors. Recently, spectroscopic evidence for spin–charge separation²⁴ was provided by the observation of different excitation branches at high energies (of the order of the conduction electron band width). But Fermi- and Luttinger-liquids differ in particular by the nature of their low-energy excitations so that it is crucial to observe consequences of this difference in a transport property, which probes the low-energy excitations and their dynamics directly. Studying the heat transport opens a new route for the investigation of low-dimensional correlated materials, including both insulating spin systems and metals. For a quantitative analysis of experiments, further theoretical work on heat transport is highly desirable. □

Methods

Large millimetre-size single crystals with typical dimensions $3 \times 0.5 \times 0.2 \text{ mm}^3$ were grown using a standard electrochemical growth technique²³. The thermal conductivity was measured with a conventional steady-state method²⁵. All results were reproducible with respect to their temperature dependence. The absolute values of k and σ scatter by up to a factor of five, presumably because of internal cracks, as commonly found in the Bechgaard salts^{7,15}. Heat flow through the thermocouple or heater leads was avoided ($< 4\%$) by using long and thin thermocouple and heater leads ($A_w/L_w \approx 30 \text{ nm}$ and $\approx 7 \text{ mm}$, respectively, where A_w is the cross-section and L_w the length of the wires). Radiation losses were minimized by the use of very small heaters with cross-sections equal to those of the samples.

Several aspects confirm the reliability of our method: (1) Radiation losses increase rapidly with increasing temperature ($\propto T^3$), yielding a positive curvature of the $k(T)$ curve, in contrast to our results especially on TM-SP. (2) The error due to radiation losses may be estimated using $k^{\text{rad}} \approx 8\epsilon\sigma_{\text{SB}}T^3L^2/A^{1/2}$, where $0 < \epsilon < 1$ is the emissivity, $\sigma_{\text{SB}} \approx 5.67 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-4}$ the Stefan–Boltzmann constant, and L and A are the sample length and cross-section, respectively. Using $\epsilon = 1$ we find an upper limit of $k^{\text{rad}}/k \leq 10\%$ at 300 K for all samples studied. For TM-SP, owing to its different geometry, we find $k^{\text{rad}}/k < 3\%$ at 300 K for both k_a and k_b , that is, k^{rad} is very small and not anisotropic. (3) We tested our set-up by measuring k of $\text{La}_{0.875}\text{Sr}_{0.125}\text{CoO}_3$ with known thermal conductivity on a sample that has similar size and shape, absolute value of k and of ϵ (judging from the black surface) as the Bechgaard salts over the entire temperature range of our measurements. The estimate of k^{rad} gives an upper limit of $k^{\text{rad}}/k \approx 8\%$ at 300 K. Experimentally, radiation losses are smaller than 1% (see inset to Fig. 3).

Received 21 March; accepted 18 June 2002; doi:10.1038/nature00913.

- Nozières, P. & Pines, D. *The Theory of Quantum Liquids* (Perseus, Cambridge, Massachusetts, 1966).
- Orenstein, J. & Millis, A. J. Advances in the physics of high-temperature superconductivity. *Science* **288**, 468–474 (2000).
- Hill, R. W., Proust, C., Taillefer, L., Fournier, P. & Greene, R. L. Breakdown of Fermi-liquid theory in a copper-oxide superconductor. *Nature* **414**, 711–715 (2001).
- Voit, J. One-dimensional Fermi liquids. *Rep. Prog. Phys.* **58**, 977–1116 (1995).
- Gogolin, A. O., Nersisyan, A. A. & Tsvelik, A. M. *Bosonization and Strongly Correlated Systems* (Cambridge Univ., Cambridge, 1998).
- Jérôme, D. & Schulz, H. J. Organic conductors and superconductors. *Adv. Phys.* **31**, 299–490 (1982).
- Ishiguro, T., Yamaji, K. & Saito, G. *Organic Superconductors* (Springer, Berlin, 1998).
- Bourbonnais, C. Organic conductors: reduced dimensionality and correlation effects. *Synth. Met.* **84**, 19–24 (1997).
- Bourbonnais, C. & Jérôme, D. One-dimensional conductors. *Phys. World* **11**(9), 41–45 (1998).

10. Bourbonnais, C. & Jérôme, D. in *Advances in Synthetic Metals, Twenty Years of Progress in Science and Technology* (eds Bernier, P., Lefrant, S. & Bidan, G.) 206–301 (Elsevier, New York, 1999).
11. Bochrath, M. *et al.* Luttinger-liquid behaviour in carbon nanotubes. *Nature* **397**, 598–601 (1999).
12. Yacoby, A., Stormer, H. L., Baldwin, K. W., Pfeiffer, L. N. & West, K. W. Magneto-transport spectroscopy on a quantum wire. *Solid State Commun.* **101**, 77–81 (1995).
13. Heeger, A. J., Kivelson, S., Schrieffer, J. R. & Su, W.-P. Solitons in conducting polymers. *Rev. Mod. Phys.* **60**, 781–850 (1988).
14. Djurek, D., Prester, M., Jérôme, D. & Bechgaard, K. Magnetic field dependent thermal conductivity in the organic superconductor (TMTSF)₂ClO₄. *J. Phys. C* **15**, L669–L674 (1982).
15. Choi, M.-Y., Chaikin, P. M. & Greene, R. L. Thermal conductivity of bis-tetramethyltetraselenafulvalen perchlorate [(TMTSF)₂ClO₄]. *Phys. Rev. B* **34**, 7727–7732 (1986).
16. Coulon, C. *et al.* A new survey of the physical properties of the (TMTTF)₂X series. Role of the counterion ordering. *J. Phys. C* **15**, 1059–1067 (1982).
17. Berman, R. *Thermal Conduction in Solids* (Clarendon, Oxford, 1976).
18. Slack, G. A. in *Solid State Physics* Vol. 34 (eds Ehrenreich, H., Seitz, F. & Turnbull, D.) 1–71 (Academic, New York, 1979).
19. Sologubenko, A. V., Gianno, K., Ott, H. R., Ammerahl, U. & Revcolevschi, A. Thermal conductivity of the hole-doped spin ladder system Sr_{14-x}Ca_xCu₂₄O₄₁. *Phys. Rev. Lett.* **84**, 2714–2717 (2000).
20. Sologubenko, A. V., Gianno, K., Ott, H. R., Vietkine, A. & Revcolevschi, A. Heat transport by lattice and spin excitations in the spin-chain compounds SrCuO₂ and Sr₂CuO₃. *Phys. Rev. B* **64**, 054412–1–054412–11 (2001).
21. Dumm, M. *et al.* Electron spin resonance studies on the organic linear-chain compound (TMTCF)₂X (C = S, Se; X = PF₆, AsF₆, ClO₄, Br). *Phys. Rev. B* **61**, 511–520 (2000).
22. Klümper, A. The spin-1/2 Heisenberg chain: thermodynamics, quantum criticality and spin-peierls exponents. *Eur. Phys. J. B* **5**, 677–685 (1998).
23. Schwarz, A. *et al.* On-chain electrodynamics of metallic (TMTSF)₂X salts: Observation of Tomonaga-Luttinger liquid response. *Phys. Rev. B* **58**, 1261–1271 (1998).
24. Claessen, R. *et al.* Spectroscopic signature of spin-charge separation in the quasi-one-dimensional organic conductor TTF-TCNQ. *Phys. Rev. Lett.* **88**, 096402–1–096402–4 (2002).
25. Zeini, B. *et al.* Thermal conductivity and thermal Hall effect in Bi- and Y-based high-T_c superconductors. *Eur. Phys. J. B* **20**, 189–208 (2001).
26. Engelsberg, S. & Varga, B. B. One-dimensional electron-photon model. *Phys. Rev.* **136**, A1582–A1589 (1964).

Acknowledgements

We thank M. Braden, C. Hess, J. Jérôme, A.P. Kampf, D.I. Khomskii, E. Müller-Hartmann, H.R. Ott and G.A. Sawatzky for discussions. This work was supported by the Deutsche Forschungsgemeinschaft and by the VolkswagenStiftung.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.F. (e-mail: freimuth@ph2.uni-koeln.de).

Sub-ångstrom resolution using aberration corrected electron optics

P. E. Batson*, N. Dellby† & O. L. Krivanek†

* IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598, USA

† Nion R&D, Kirkland, Washington 98033, USA

Following the invention of electron optics during the 1930s, lens aberrations have limited the achievable spatial resolution to about 50 times the wavelength of the imaging electrons¹. This situation is similar to that faced by Leeuwenhoek in the seventeenth century, whose work to improve the quality of glass lenses led directly to his discovery of the ubiquitous “animalcules” in canal water, the first hints of the cellular basis of life. The electron optical aberration problem was well understood from the start, but more than 60 years elapsed before a practical correction scheme for electron microscopy was demonstrated², and even then the remaining chromatic aberrations still limited the resolution. We report here the implementation of a computer-controlled aberration correction system in a scanning transmission electron microscope³, which is less sensitive to chromatic aberrations.

Using this approach, we achieve an electron probe smaller than 1 Å. This performance, about 20 times the electron wavelength at 120 keV energy, allows dynamic imaging of single atoms, clusters of a few atoms, and single atomic layer ‘rafts’ of atoms coexisting with Au islands on a carbon substrate. This technique should also allow atomic column imaging of semiconductors, for detection of single dopant atoms, using an electron beam with energy below the damage threshold for silicon.

Over 50 years ago, Scherzer¹ recognized that round electron lenses would suffer from spherical aberration, limiting spatial resolution in an electron microscope to about 2 Å for electrons of energy 100–200 keV. This is unfortunate, because at this resolution the atomic structure of most crystalline materials can be imaged in only one projection. Improving the resolution to 0.5 Å would increase the number of available crystal projections to more than twelve, allowing three-dimensional imaging of atomic structure. This problem is illustrated by the recent detection of individual dopant atoms within bulk semiconductors⁴, where structural characterization of the local environment of the dopant atoms was complicated by the 1.6-Å probe size, slightly larger than the 1.36-Å distance between atoms in the {110} projection of Si.

Scherzer also pointed out three possible ways to correct the limiting aberrations: (1) use of non-round lenses; (2) use of lenses

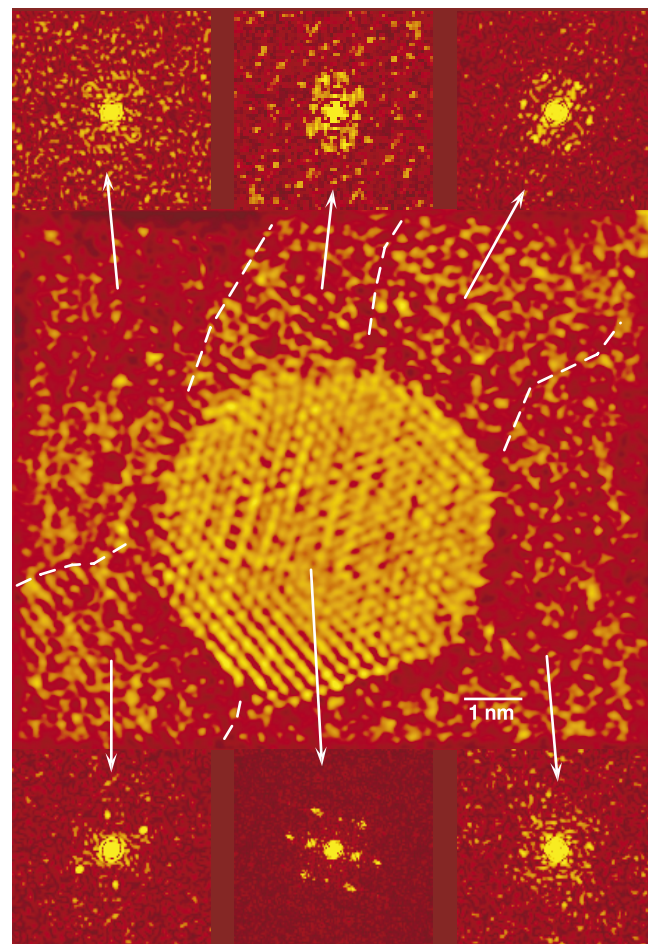


Figure 1 Atomic resolution image of a Au island on an amorphous carbon substrate. Surrounding the island are ‘rafts’ of single atomic layers of Au. Further away, small clusters and single atoms of Au are present. Diffraction patterns from various regions surrounding the island show that the rafts are ordered in various structures adjacent to the built-up islands.

with charge on axis (including mirrors); and (3) use of time-varying fields. Since his work, many attempts have been made to develop practical aberration correctors for the transmission electron microscope (TEM), culminating in the successful work of Haider *et al.* in the 1990s (ref. 2), which was the first to demonstrate a resolution improvement in a particular instrument. Difficulties remain, however, because the TEM requires a faithful treatment of electron paths which are far from the optic axis, necessitating chromatic aberration correction or addition of an electron monochromator, a task which is not yet complete.

In the late 1960s, an alternative to the TEM imaging geometry was introduced by Crewe *et al.*³. They used similar optics to produce a very small electron probe that was scanned in a raster over the area of interest. This geometry, called the scanning transmission electron microscope (STEM), has become an important tool for quantitative microscopy because many types of analytical signals can be used to produce an image of the scanned area. In particular, annular dark field (ADF) imaging, using high-angle elastic scattering which occurs near individual atoms⁵, has emerged as the primary imaging technique. ADF STEM imaging enjoys several advantages over conventional TEM imaging: (1) the spatial resolution is somewhat better; (2) it is sensitive to the atomic number of the imaged atoms; and (3) it provides a positive definite transfer of specimen spatial frequencies, allowing a direct interpretation of results with fewer ambiguities. As the resolution of instruments has improved, ADF imaging has made important contributions to the imaging of individual columns of atoms in crystals^{6,7} and, more recently, to the single dopant atom detection within the bulk⁴. The STEM is also compatible with other analytical techniques, including electron energy loss spectroscopy (EELS)⁸. This technique has proved especially powerful, producing atomic resolution information about atomic species, bonding environment and local electronic structure^{9–11}.

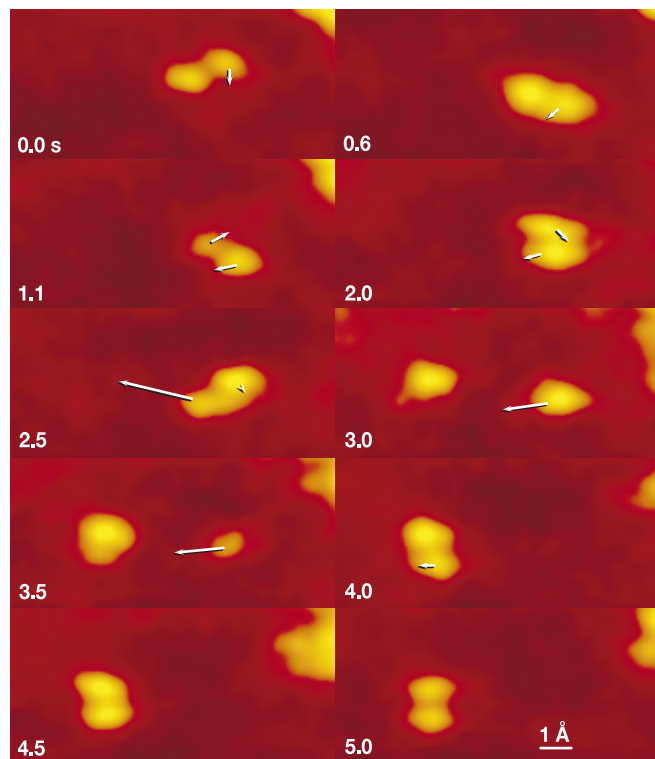


Figure 2 Selected frames from a 10 frames s^{-1} acquisition of the interaction of two Au atoms. The interaction is clearly attractive beyond about 1.5 Å and repulsive at smaller distances. Motion occurs in fast jumps, indicated by the arrows, with relatively long periods at rest. At 3.5 s in this sequence, a jump appears to occur within one frame.

The instrument used in this work is based on the VG Microscopes HB501 STEM with the addition of a quadrupole–octupole aberration corrector¹². Before the incorporation of the corrector, the IBM STEM had been used extensively for ADF imaging and spatially resolved EELS of defects and interfaces in semiconductors⁹. At the 1.9-Å resolution level, it yielded tantalizing, but incomplete information about the atomic structure of important defects such as the dissociated 60° dislocation in Si-Ge/Si structures¹³. The corrector comprises seven major elements: four quadrupoles separated by three octupoles. In addition, minor windings include dipoles and quadrupoles at each stage to allow the control of parasitic aberrations. In all, 35 windings are excited separately using 0.3 p.p.m. stability, computer-controlled current supplies. All axial aberrations up to fifth order in angle are determined by software which analyses electron shadow images (Ronchigrams¹⁴) obtained by a television camera located in the far field behind the specimen. The software then adjusts the corrector to compensate for aberrations up to fourth order. The system spherical aberration (third order) is finally adjusted to about $-50 \mu m$ to optimally oppose the effects of fifth order aberrations, which are not corrected in the present instru-

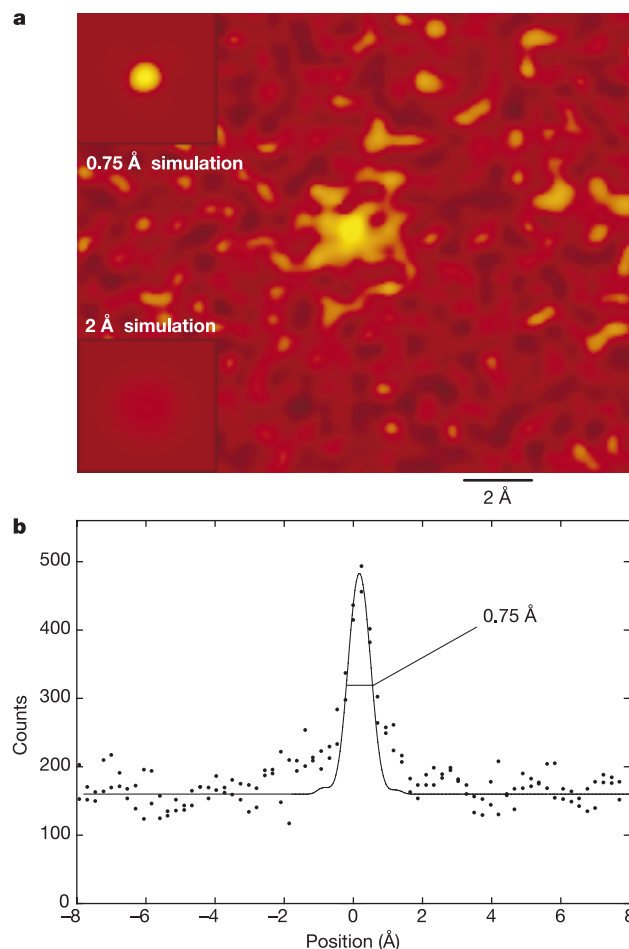


Figure 3 An analysis of the image size of a single Au atom. **a**, Image of a single Au atom on a carbon support compared with simulations for the corrected (upper inset) and uncorrected (lower inset) instrument. The upper inset, showing 0.75 Å resolution, is a full calculation for the ADF STEM using the Au projected potential, the measured aberration coefficients, and a source size of 0.3 Å, estimated using the source brightness (about $10^9 \text{ Å cm}^{-2} \text{ sr}^{-1}$), the measured probe current (40 pA) and the 50 mrad full illumination angle. The atom would not be visible at the 1.9 Å resolution, shown in the lower left, obtained for a spherical aberration coefficient of 1.2 mm in the uncorrected instrument. **b**, Line scan data, obtained from the image, are compared with the 0.75 Å probe calculation.

ment. Details are described elsewhere¹².

Preliminary imaging of 100-Å Au islands on amorphous carbon have yielded many images similar to that shown in Fig. 1. We see, in addition to the islands, a hierarchy of structures: flat areas with no apparent Au atoms, single atoms, clusters of 2–3 atoms, and 20–50-Å-wide single layer rafts of Au atoms. We identify them as Au layers on the basis of the similarity of their image contrast with that of single atoms produced during dissolution of small Au islands. The monolayer rafts appear to be transition regions between the islands and the more sparsely occupied areas further away, and so we speculate that they may be important to the exchange of atoms between the island and the substrate during growth¹⁵. Finally, the single atom contrast was high enough to follow the motion of atoms under the electron beam, during dissolution of islands and during rapid reformation of islands from dilute groups of atoms.

Figure 2 summarizes a 10 frames s⁻¹ acquisition of the interaction between two Au atoms at 300 K on the carbon surface. An attractive interaction is clearly visible when the separation is greater than about 1.3–1.5 Å, but at no time do the Au atoms approach closer than this. In this sequence, the two atoms approach, circle one another, separate by about 4 Å and then rejoin. Several things need to be investigated to understand this process: (1) the substrate is

complex, possibly providing low-energy channels for atomic motion; (2) the electron beam imparts a finite impulse to the atoms during the imaging process, but the magnitude of this impulse relative to thermal fluctuations needs to be evaluated; (3) higher-speed acquisition may provide information about the diffusion velocity in this situation.

Fourier transform power spectra for the images such as Fig. 1 show spatial frequencies as small as 1.17 Å, but this does not give us a good measure of the probe size for several reasons. For instance, small crystal tilts in the islands can lead to blurring of the atom columns. Errors in acquisition of the image intensity, particularly in the low-intensity region between atom columns in crystals, can lead to nonphysical spatial periodicities (Yu, Z. and Silcox, J., personal communication). Therefore, we have used images of single Au atoms to evaluate the probe size. This evaluation is summarized in Fig. 3a, where we show an image of a single Au atom compared with calculated images for the corrected and uncorrected instrument. We estimate the probe shape on the basis of the measured system aberration parameters, obtained from an analysis of the electron shadow map at the time of the experiment, and the finite source size, obtaining a corrected probe size of 0.74 Å and an uncorrected size of 1.9 Å at 120 keV. Using a full ADF STEM image calculation for the Au atom¹⁶, we obtain a corrected image width of 0.75 Å. The visual correspondence is good, but more interestingly, it is obvious that the 1.9 Å resolution is not good enough even to detect single Au atoms in the presence of noise from the carbon support. In Fig. 3b, we compare the 0.75 Å profile with data obtained from the atom image. Although the statistical quality for the experiment is not sufficient to obtain much detail regarding the shape of the image, certainly the width is consistent with the calculated width.

Figure 4 shows results for a Ge₃₀Si₇₀ alloy with and without aberration correction. The shadow maps in Fig. 4a and b show regions of minimum contrast at the centre defining an area of constant electron phase, appropriate for use in forming a small probe. The larger available area using aberration correction produces the smaller probe, consistent with the Heisenberg uncertainty relation. Figure 4c and d shows the resulting crystal image. Markers highlight the well-known 1.36 Å 'dumb-bell' structure formed by the near-neighbour atom columns. In Fig. 4e and f the spatial power spectra for the images is shown. In the uncorrected result, a few spacings out to 1.92 Å are present, illustrating the best previous performance. In the corrected result, many spacings are present, including {551} planes with 0.76 Å spacings. This is a great improvement over that achieved in the past in essentially the same instrument¹². We caution, however, that care needs to be applied in interpreting this image because high spatial frequencies can be induced by nonlinearities in the image acquisition (Yu, Z. and Silcox, J., personal communication). This has not been observed in the past at 1.9 Å resolution, but does appear to be present at some level in the new results. Even so, the dramatic contrasts between the images are obvious.

To our knowledge, no other electron microscope has attained sub-ångstrom directly interpretable resolution in a single image. Lattice spacings smaller than 1 Å have been observed using high beam energy¹⁷, and sub-ångstrom resolution has been obtained by *a posteriori* processing¹⁸. However, the electron phases in the raw data in such work are always severely distorted, and no structural conclusions are possible without extensive computer processing. Moreover, the acquisition for this approach can exceed one hour, thus exacerbating beam damage, particularly for light elements. Structural resolution much better than 1 Å can also be obtained from electron diffraction data¹⁹. However, this approach is limited to periodic structures, and also requires extensive computer processing.

This successful aberration correction is linked to recent technical advances: (1) computation of electron optical parameters is now

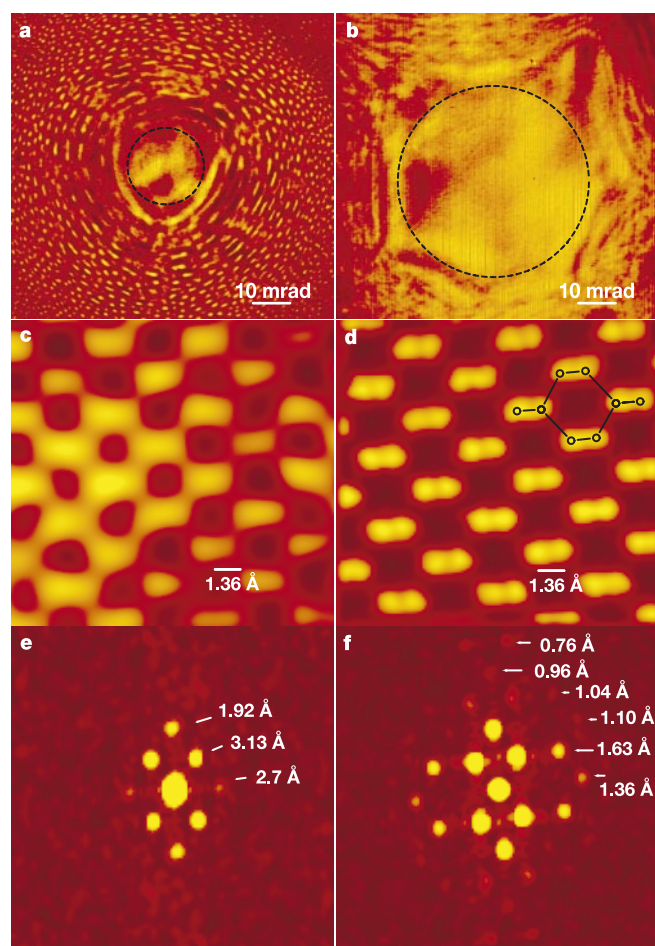


Figure 4 Summary of results for the {110} projection of Ge₃₀Si₇₀. **a, b**, Ronchigram shadow maps for uncorrected and corrected imaging. Circles indicate the range of angles used for probe formation. **c, d**, Images of the {110} projection of the crystal obtained using the conditions in **a** and **b**. Lines and circles indicate the projected atomic structure. The image on the left is the result of a sum of several 0.5-s acquisitions, while that on the right is a sum of several sections cut from a single large area image obtained in 0.5-s, using a beam current of about 40 pA. **e, f**, The Fourier power spectra for the two examples in **c** and **d**. Spatial frequencies are marked.

possible for nonrotationally symmetric systems, allowing practical designs to be simulated with high accuracy; (2) mechanical fabrication tolerances have advanced materially in the past 15 years; (3) high stability electronic components have become available in the past 10 years, allowing the packaging of many, very high stability, computer-controlled power supplies in a small space; and (4) high-speed small computers are now available for real-time processing of the shadow map Rochigram data to obtain aberration parameters.

We believe aberration correction technology combined with the inherent power of the STEM instrument opens the way for lower voltage, smaller, 'smarter', more easily managed instruments that will be capable of routinely imaging and analysing materials at sub-ångström resolution, using several different crystal orientations. Specialized instruments, perhaps using some submillimetre-sized electrostatic imaging elements, may reach the diffraction limit for the imaging electrons. Multiple corrector systems will allow aberration control of both probe size and detector field of view, in a manner somewhat similar to confocal light microscopy. Finally, detailed control of the amplitude and phase of the incident and scattered electron wavefunction will be possible, allowing specific specimen inelastic transitions to be accessed²⁰. Thus aberration correction marks a shift towards precise optical control that will allow routine atomic level characterization of defects and interfaces within bulk materials. □

Received 5 April; accepted 4 July 2002; doi:10.1038/nature00972.

1. Scherzer, O. The theoretical resolution limit of the electron microscope. *J. Appl. Phys.* **20**, 20–29 (1949).
2. Haider, M., Uhlemann, S., Schwan, E., Kabius, B. & Urban, K. Electron microscopy image enhanced. *Nature* **392**, 768–769 (1998).
3. Crewe, A. V., Isaacson, M. & Johnson, D. A simple scanning electron microscope. *Rev. Sci. Instrum.* **40**, 241–246 (1969).
4. Voyles, P. M., Muller, D. A., Grazul, J. L., Citrin, P. H. & Gossmann, H.-J. L. Atomic-scale imaging of individual dopant atoms and clusters in highly n-doped bulk Si. *Nature* **416**, 826–829 (2002).
5. Crewe, A. V., Wall, J. & Langmore, J. Visibility of a single atom. *Science* **168**, 1338–1340 (1970).
6. Pennycook, S. J. & Boatner, L. A. Chemically sensitive structure-imaging with a scanning transmission electron microscope. *Nature* **336**, 565–567 (1988).
7. Xu, P., Kirkland, E. J., Silcox, J. & Keyse, R. High resolution imaging of silicon (111) using a 100 KeV STEM. *Ultramicroscopy* **32**, 93–102 (1990).
8. Crewe, A. V., Isaacson, M. & Johnson, D. A high resolution electron spectrometer for use in transmission scanning microscopy. *Rev. Sci. Instrum.* **42**, 411–420 (1971).
9. Batson, P. E. Simultaneous STEM imaging and electron-energy-loss-spectroscopy with atomic-column sensitivity. *Nature* **366**, 727–728 (1993).
10. Muller, D. A., Tzou, Y., Raj, R. & Silcox, J. High resolution EELS at grain boundaries. *Nature* **366**, 725–727 (1993).
11. Browning, N. D., Chisholm, M. F. & Pennycook, S. J. Atomic resolution analysis using a scanning transmission electron microscope. *Nature* **366**, 143–146 (1993).
12. Dellby, N., Krivanek, O. L., Nellist, P. D., Batson, P. E. & Lupini, A. R. Progress in aberration-corrected scanning transmission electron microscopy. *J. Electron. Microsc.* **50**, 177–185 (2001).
13. Batson, P. E. Structural and electronic characterization of a dissociated 60° dislocation in GeSi. *Phys. Rev. B* **61**, 16633–16641 (2000).
14. Ronchi, V. Forty years of history of a grating interferometer. *Appl. Opt.* **3**, 437–450 (1964).
15. Zhang, Z. Y. & Lagally, M. G. Atomistic processes in the early stages of thin film growth. *Science* **276**, 377–383 (1997).
16. Kirkland, E. J. *Advanced Computing in Electron Microscopy* (Plenum, New York, 1998).
17. Kawasaki, T. et al. Development of a 1 MV field emission transmission electron microscope. *J. Electron. Microsc.* **49**, 711–718 (2000).
18. O'Keefe, M. A. et al. Sub-ångström high resolution transmission electron microscopy at 300 keV. *Ultramicroscopy* **89**, 215–241 (2001).
19. Zuo, J., Kim, Y., O'Keefe, M. & Spence, J. C. H. Direct observation of d holes and Cu-Cu bonding in Cu₂O. *Nature* **401**, 49–56 (1999).
20. Batson, P. E. Symmetry selected electron energy loss scattering in diamond. *Phys. Rev. Lett.* **70**, 1822–1825 (1993).

Acknowledgements

O.L.K. and N.D. acknowledge partial support for this project from the IBM Corporation.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.E.B. (e-mail: batson@us.ibm.com).

Experimental evidence for sub-3-fs charge transfer from an aromatic adsorbate to a semiconductor

Joachim Schnadt*, Paul A. Brühwiler*, Luc Patthey†, James N. O'Shea*, Sven Södergren*, Michael Odelius‡, Rajeev Ahuja*, Olof Karis*, Margit Bässler§, Petter Persson*, Hans Siegbahn*, S. Lunell* & Nils Mårtensson§*

* Department of Physics, Uppsala University, Box 530; ‡ Department of Physical Chemistry, Uppsala University, Box 532, 75121 Uppsala, Sweden

† Swiss Light Source, Paul-Scherrer-Institut, 5232 Villigen-PSI, Switzerland

§ MAX-Lab, University of Lund, Box 118, 22100 Lund, Sweden

The ultrafast timescale of electron transfer processes is crucial to their role in many biological systems and technological devices. In dye-sensitized solar cells^{1–4}, the electron transfer from photo-excited dye molecules to nanostructured semiconductor substrates needs to be sufficiently fast to compete effectively against loss processes and thus achieve high solar energy conversion efficiencies⁴. Time-resolved laser techniques indicate an upper limit of 20 to 100 femtoseconds^{5–9} for the time needed to inject an electron from a dye into a semiconductor, which corresponds to the timescale on which competing processes such as charge redistribution^{10,11} and intramolecular thermalization of excited states^{12–14} occur. Here we use resonant photoemission spectroscopy, which has previously been used to monitor electron transfer in simple systems with an order-of-magnitude improvement in time resolution^{15,16}, to show that electron transfer from an aromatic adsorbate to a TiO₂ semiconductor surface can occur in less than 3 fs. These results directly confirm that electronic coupling of the aromatic molecule to its substrate is sufficiently strong to suppress competing processes¹⁷.

The most efficient photoelectrochemical solar cells are currently based on a nanostructured TiO₂ electrode³. The electrode is sensitized to visible light by adsorption of a dye. One of the most promising dyes is "N3"¹⁸ or (dcb)₂Ru(NCS)₂ (dcb = 4,4'-dicarboxy-2,2'-bipyridine, also termed bi-isonicotinic acid), which binds to the nanostructured electrode through one¹⁹ or two¹ dcb

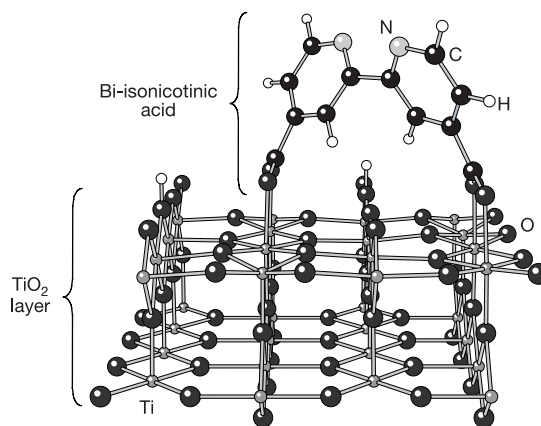


Figure 1 Geometry of the adsorption structure of bi-isonicotinic acid on TiO₂(110). The bonding is on-top for each oxygen atom (2M-bidentate) in both carboxyl groups, with the dissociated hydrogen atoms (small white spheres) assumed to bond to nearby surface oxygen atoms. The pyridine rings are tilted in opposite directions around the inter-ring bond (large white spheres, nitrogen; large black spheres, carbon). The structural model was further refined for the present study using first-principles plane-wave density functional calculations and performing a full optimization of the adsorbate and the surface TiO₂ layer (small white spheres, titanium; large black spheres, oxygen).

ligands. For this dye the above upper limits on the electron transfer time have been measured^{5–9}. As a model interface for this important system we have chosen a monolayer of dcb adsorbed on a rutile TiO₂(110) surface²⁰. The structure of the interface is illustrated in Fig. 1. The carboxylic groups bind to the substrate in a 2M-bidentate fashion, that is, with each of the carboxylic oxygen atoms attached to one substrate Ti-atom. This geometry is similar to the geometry surmised for bi-isonicotinic acid on anatase TiO₂(101)²¹ and on anatase TiO₂ nanoparticles grown *in situ* (J.S. *et al.*, manuscript in preparation) as well as to one of two suggested geometries of the original cell^{1,19}. The bandgap of rutile TiO₂ (3.05 eV)²² is close to that of the nanostructured anatase used in the photoelectrochemical cell (3.20 eV)²³. The high structural quality of a single-crystal surface in ultrahigh vacuum allows a detailed study of the electronic structure of the interface.

The experimental techniques include photoemission spectroscopy (PES), X-ray absorption spectroscopy (XAS), and resonant photoemission spectroscopy (RPES), which are described in the Methods section and in Fig. 2. We employ PES and XAS to assess the electronic density-of-states. RPES probes the interfacial electron transfer dynamics in the low-fs and sub-fs time regimes by monitoring those decay events of the core-excited state that involve the excited electron. A transfer of the excited electron competes with the decay process that in our case takes place on a 6-fs timescale and thus modifies the RPES intensity. The RPES intensities can be compared to the XAS intensities to give electron transfer times (see Methods section for details). In principle, the same information could be derived from a measurement of the XAS lifetime-induced linewidth. However, unresolved vibrational broadening, typical for solid samples, renders such a direct measurement impossible. The periodic pseudopotential calculations provide supporting information about the geometric and electronic structure of the system, and of the interfacial electronic coupling.

The electronic structure of the interface is presented in Fig. 3. The bandgap of the substrate is located between 5 and 8 eV binding energy. We have studied the first three unoccupied states of the adsorbate. The lowest unoccupied molecular orbital (LUMO) is situated almost completely within the bandgap. This is an effect of the core hole (Fig. 2b) produced in XAS, which leads to energy shifts to higher binding energies²⁴, similar to the case of valence excitations²⁵. The other two orbitals (LUMO+1 and LUMO+2), in contrast, lie within the substrate conduction band. The calculations

confirm these fundamental experimental results. These findings imply that in the presence of the core hole an electron residing in the LUMO cannot be transferred into the substrate, while for an electron in the LUMO+1 and LUMO+2 a transfer is energetically feasible. We now use RPES to show that the latter cases indeed lead to a very rapid electron delocalization into the substrate.

Figure 4 shows the RPES and contains the XAS from Fig. 3 for comparison. The large peak at 398.6 eV in the RPES shows that the excited electron promoted to the LUMO is localized on the molecule, in agreement with the results for the density-of-states. In contrast, the signals observed in the XAS for the LUMO+1 and LUMO+2 are completely quenched in the RPES, indicating ultra-fast delocalization of the excited electron into the substrate. In Fig. 4 we indicate the noise level N of the RPES and the XAS signal intensity S . N is at most 10% of the XAS intensity, that is, $N/S < 0.1$. A possible RPES signal for the LUMO+2 hidden by the noise could thus at most amount to 10% of the XAS intensity S , that is $I_{\text{RPES}}/I_{\text{XAS}} < 0.1$ in terms of the intensities in the RPES and the XAS. Using the equation given in the Methods section—that is, $\tau_T = \tau_C(I_{\text{RPES}}/I_{\text{XAS}})/[C - (I_{\text{RPES}}/I_{\text{XAS}})]$ where $\tau_C = 6$ fs and $C \approx 1/3$ determined from the data in the Supplementary Information—this result corresponds to an upper limit on the electron transfer time, $\tau_T < 2.5$ fs, much faster than previous studies could resolve. For the LUMO+1 the noise level is slightly higher, leading to an upper limit $\tau_T < 3$ fs, which with little loss of precision we adopt for both the LUMO+1 and LUMO+2 states.

This result is semiquantitatively explained by the calculations, which show quite broad LUMO+1 and LUMO+2 bands (compare Fig. 3), with a width of the order of 1.5 eV. Interpreted as bandwidths, and assuming that no vibrational or other broadening effects exist, this corresponds to electron transfer times much shorter than 3 fs. Closer agreement on the resonance energy has been achieved for a simpler system²⁶, and a similar level of agreement for systems of the present complexity is a challenge for the near future. A key aspect of the present study is that the soft-X-ray excitation frequencies used here are much higher than the fastest possible electron transfer rate. Application of these short electron transfer times to the optical excitations relevant to solar cell applications requires care, because the timescales of both processes are comparable in such cases. Because of this, optical excitation and

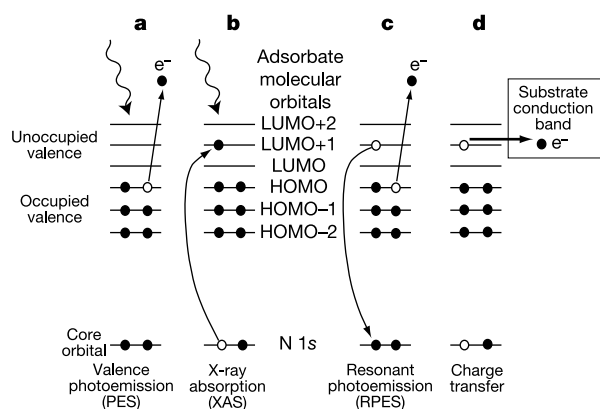


Figure 2 Relevant processes. **a–c**, PES, XAS and RPES applied to a molecule. The electronic states are indicated by lines. PES measures the occupied and XAS the local unoccupied density-of-states. RPES monitors the decay of the excited state. **d**, If the molecular orbital couples to the substrate density-of-states an electron transfer may be possible that competes with RPES. HOMO, highest occupied molecular orbital; LUMO, lowest unoccupied molecular orbital. Here, an electron transfer is energetically possible from LUMO+1 and LUMO+2, but not from the LUMO. To the extent that the depicted electron transfer occurs, the RPES is reduced in spectral intensity.

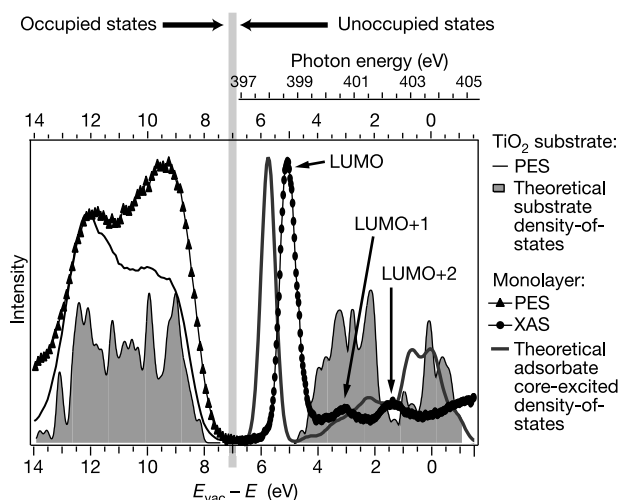


Figure 3 Density-of-states for bi-isonicotinic acid on TiO₂(110): experimental monolayer density-of-states, experimental clean substrate occupied density-of-states, theoretical substrate density-of-states, and XAS calculation. Besides the common energy scale for XAS and PES¹⁶, the photon energy scale for the experimental XAS is given. The LUMO of the excited system lies within the bandgap, but the higher-lying states overlap the conduction band density-of-states. Electronic overlap is necessary for any electron transfer to take place.

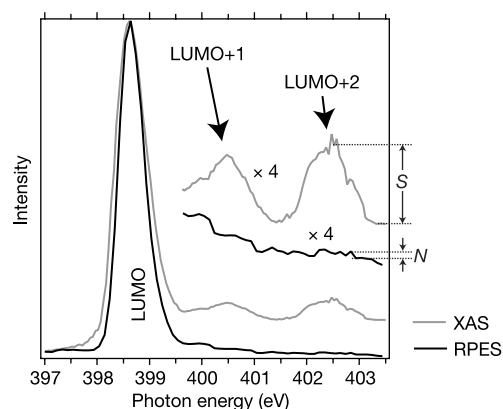


Figure 4 Comparison of RPES and XAS for a bi-isonicotinic acid monolayer. Excitation to the LUMO leads to a large intensity in the RPES owing to strong resonant photoemission, indicating localization of the excited electron on the molecule. For excitation to LUMO+1 and LUMO+2 no RPES-intensity is observed, which implies rapid excited electron

transfer of the electron cannot be disentangled. Hence it is more appropriate to discuss the adsorbate–substrate interaction in terms of coupling strength, that is, in terms of an electron transfer bandwidth (for example, ref. 27), rather than in terms of a well-defined electron transfer time. From the measured electron transfer time of less than 3 fs for a core excitation we can conclude that the system is characterized by strong interfacial electronic coupling. Together with previous calculations²⁴ the present results show that this extremely rapid transfer relies on the overlap of ring, carboxylic group and substrate electronic states. The weak-coupling Marcus theory commonly used to model electron transfer reactions (for reviews and discussion see, for example, refs 2, 4 and 17) is inappropriate to describe the electronic coupling at these interfaces. This is confirmed by the observation that the electron transfer rate is not limited by the timescale of vibrational motion in this case, although vibrations are excited in the N 1s as in optical transitions, here easily observed as the dominant contribution to the width of the LUMO resonance. Similar vibronic coupling is expected for the LUMO+1 and LUMO+2 states. This observation is consistent with at least three models^{17,28,29}. Furthermore, the electron transfer takes place on a much faster timescale than charge redistribution^{10,11} and thermalization^{12–14} in N3 and similar systems.

Aromatic rings such as those in dcb are well known to facilitate charge delocalization. Here we establish that the carboxylic acid group, which forms a covalent bond to the oxide substrate, completes an efficient conduit of electric current from the N atoms into the substrate; the coupling is comparable to the strongest observed for CO on metal surfaces¹⁶. The large distance introduced by the carboxylic spacer group could be expected to lengthen the electron transfer time²⁸. The present results show in contrast that the ligand presents a almost negligible barrier, helping to rationalize the observed ultrafast electron transfer for the N3 dye itself^{5–8}. The quantitative application of the present results to the N3 case awaits the ultimate characterization of the N3 bonding geometry. However, the delocalized π -bands will tend to compensate for a single¹ as opposed to a double¹⁹ carboxylate bond, which is also apparent in calculated wavefunctions for the present case²⁴. The knowledge gained here should be valuable in strategies for optimizing the charge transfer characteristics of future photovoltaic systems. □

Methods

Photoemission spectroscopy

PES maps the energies of the occupied electronic states of a sample, that is, it images the occupied electronic density-of-states. In the case presented here an electron from an occupied valence state is excited by absorption of a photon with a well-defined photon energy $h\nu$ into the vacuum (Fig. 2a) to kinetic energies E_K of approximately 100 eV. A so-called electron binding energy E_B in the adsorbate–substrate system is derived via

delocalization. For the LUMO+1 and LUMO+2, we estimated the maximum possible RPES signal consistent with the noise level N , and compared it to the XAS signal S , determining the upper limits in the text. The XAS has a 12:1 signal-to-noise ratio at the LUMO+2.

$E_B = h\nu - E_K$ and can often be mapped onto electronic density-of-states ground state calculations. Here we call the adsorbate state at lowest binding energy the HOMO (highest occupied molecular orbital), the next one HOMO–1, and so on. The energy scale has its origin at the vacuum level E_{vac} (ref. 16). Clean crystal and monolayer PES spectra were mutually aligned using the O 2s level.

X-ray absorption spectroscopy

XAS maps the unoccupied electronic density-of-states. In the present case (Fig. 2b), absorption of a photon with well-defined energy $h\nu$ promotes an N 1s core electron to a bound unoccupied state with p -character. We term this electron the excited electron. By varying the photon energy different unoccupied states can be reached, and the small spatial extent of the 1s orbital ensures that adsorbate states are probed. The first observed state is referred to as the lowest unoccupied molecular orbital (LUMO), followed by the LUMO+1, and so on. The process of removing an electron from a core state, that is, the creation of a core-hole, can quantitatively be viewed as an increase of the positive nuclear charge by one unit^{16,26}. Owing to this the observed energies are lower than the calculated ground state energies, especially for the LUMO. The XAS was placed¹⁶ onto the same energy scale as PES.

Density-of-states and XAS calculations

The theoretical curve for the density-of-states is a periodic pseudopotential calculation with gradient corrected functionals of the total ground state density-of-states and was performed using version 3.3 of the CPMD code (for in more detail see <http://www.fz-juelich.de/nic-series/Volume3/marx.pdf>). It has been aligned to the PES data of clean TiO₂(110), with the gap artificially adjusted to the measured optical one²². The XAS has been modelled by calculating the p -like density-of-states with an excited N 1s electron, which is directly comparable to the experimental data²⁴.

Resonant photoemission spectroscopy

In XAS the system is left in an excited state with an additional electron in a bound unoccupied level and a vacancy (a missing electron) in a core (N 1s) level. This core-excited system will decay exponentially via Auger processes (99.9%), with a time constant given by the core-hole lifetime τ_C (6 fs for N 1s). As shown in Fig. 2c, a valence electron (or the excited electron itself) fills the core hole, while another valence electron (or the excited electron) takes up the energy released in that process and is ejected into the vacuum and measured. RPES monitors only decay processes in which the excited electron takes part. These processes are termed resonant photoemission. Because the remaining system is left in the same state as for PES (compare Fig. 2a), RPES measurements are obtained by integrating the valence PES containing the resonant photoemission signal for the photon energies studied in XAS, that is, one probes the effect of placing the excited electron in different unoccupied states. The RPES intensity is then compared to the XAS intensity for each unoccupied state, and results in a particular state-dependent ratio that we call C for the case of an isolated molecule (in our case, $C \approx 1/3$, see Supplementary Information). C describes the sensitivity of RPES compared to XAS. If the particular unoccupied orbital to which the excited electron was promoted is coupled to unoccupied substrate states, the excited electron delocalizes into the substrate in an exponential process (to first order) with the characteristic electron transfer time τ_T . Thus the transfer process competes with the decay process characterized by τ_C and monitored in RPES (Fig. 2c and d). If the two timescales are comparable the RPES intensity serves as a fingerprint for the extent of excited electron localization. In principle, this enables RPES to probe the electron transfer dynamics within roughly one order of magnitude (up or down) of the core-hole lifetime^{15,16}. The transfer time of the excited electron into the substrate can be estimated from the XAS and RPES intensities I_{XAS} and I_{RPES} for the particular unoccupied state from¹⁶ $\tau_T = \tau_C(I_{RPES}/I_{XAS})/[C - (I_{RPES}/I_{XAS})]$. For very fast transfers the RPES vanishes completely for the particular state. This simple model can be applied because the XAS excitation occurs on the 10-attosecond (10^{-17} s) timescale, far shorter than attainable electron transfer times and also far faster than optical excitations. The RPES, the best of several runs, contains a linear, slowly decreasing, direct photoemission background.

Received 11 April 2001; accepted 1 July 2002; doi:10.1038/nature00952.

- Hagfeldt, A. & Grätzel, M. Molecular photovoltaics. *Acc. Chem. Res.* **33**, 269–277 (2000).
- Vlček, A. Jr The life and times of excited states of organometallic and coordination compounds. *Coord. Chem. Rev.* **200–202**, 933–977 (2000).
- Grätzel, M. Photoelectrochemical cells. *Nature* **414**, 338–344 (2001).
- Miller, R. J. D., McLendon, G. L., Nozik, A. J., Schmickler, W. & Willig, F. *Surface Electron Transfer Processes* (VCH Publishers, New York, 1995).
- Tachibana, Y., Moser, J. E., Grätzel, M., Klug, D. R. & Durrant, J. R. Subpicosecond interfacial charge separation in dye-sensitized nanocrystalline titanium dioxide films. *J. Phys. Chem.* **100**, 20056–20062 (1996).
- Hannappel, T., Burfeindt, B., Storck, W. & Willig, F. Measurement of ultrafast photoinduced electron transfer from chemically anchored Ru-dye molecules into empty electronic states in a colloidal anatase TiO₂ film. *J. Phys. Chem. B* **101**, 6799–6802 (1997).
- Asbury, J. B., Hao, E., Wang, Y., Ghosh, H. N. & Lian, T. Ultrafast electron transfer dynamics from molecular adsorbates to semiconductor nanocrystalline thin films. *J. Phys. Chem. B* **105**, 4545–4557 (2001).
- Heimer, T. A., Heilweil, E. J., Bignozzi, C. A. & Meyer, G. J. Electron injection, recombination, and halide oxidation dynamics at dye-sensitized metal oxide interfaces. *J. Phys. Chem. A* **104**, 4256–4262 (2000).
- Benkő, G., Kallioinen, J., Korppi-Tommola, J. E. I., Yartsev, A. P. & Sundström, V. Photoinduced ultrafast dye-to semiconductor electron injection from nonthermalized and thermalized donor states. *J. Am. Chem. Soc.* **124**, 489–493 (2002).
- Yeh, A. T., Shank, C. V. & McCusker, J. K. Ultrafast electron localization dynamics following photo-induced charge transfer. *Science* **289**, 935–938 (2000).
- Tachibana, Y., Haque, S. A., Mercer, I. P., Durrant, J. R. & Klug, D. R. Electron injection and recombination in dye sensitized nanocrystalline titanium dioxide films: A comparison of ruthenium bipyridyl and porphyrin sensitizer dyes. *J. Phys. Chem. B* **104**, 1198–1205 (2000).
- Damrauer, N. H. *et al.* Femtosecond dynamics of excited-state evolution in [Ru(bpy)₃]²⁺. *Science* **275**, 54–57 (1997).
- Blanchet, V., Zgierski, M. Z., Seideman, T. & Stolow, A. Discerning vibronic molecular dynamics using time-resolved photoelectron spectroscopy. *Nature* **401**, 52–54 (1999).
- Willig, F., Zimmermann, C., Ramakrishna, S. & Storck, W. Ultrafast dynamics of light-induced electron injection from a molecular donor into the wide conduction band of a semiconductor as acceptor. *Electrochim. Acta* **45**, 4565–4575 (2000).
- Wurth, W. & Menzel, D. Ultrafast electron dynamics at surfaces probed by resonant Auger spectroscopy. *Chem. Phys.* **251**, 141–149 (2000).
- Brühwiler, P. A., Karis, O. & Mårtensson, N. Charge transfer dynamics studied using resonant core spectroscopies. *Rev. Mod. Phys.* **74**, 703–740 (2002).
- Lanzafame, J. M., Palese, S., Wang, D., Miller, R. J. D. & Muentner, A. A. Ultrafast nonlinear optical studies of surface reaction dynamics: Mapping the electron trajectory. *J. Phys. Chem.* **98**, 11020–11033 (1994).
- Nazeeruddin, M. K. *et al.* Conversion of light to electricity by cis-X₂bis(2,2′-bipyridyl-4,4′-dicarboxylate)ruthenium(II) charge transfer sensitizers (X = Cl[−], Br[−], I[−], CN[−], and SCN[−]) on nanocrystalline TiO₂ electrodes. *J. Am. Chem. Soc.* **115**, 6382–6390 (1993).
- Rensmo, H. *et al.* XPS studies of Ru-polypyridine complexes for solar cell applications. *J. Chem. Phys.* **111**, 2744–2750 (1999).
- Patthey, L. *et al.* Adsorption of bi-isonicotinic acid on rutile TiO₂(110). *J. Chem. Phys.* **110**, 5913–5918 (1999).
- Persson, P. & Lunell, S. Binding of bi-isonicotinic acid to anatase TiO₂(101). *Solar Energy Mater. Solar Cells* **63**, 139–148 (2000).
- Cronmeyer, D. C. Electrical and optical properties of rutile single crystals. *Phys. Rev.* **87**, 876–886 (1952).
- O'Regan, B. & Grätzel, M. A low-cost, high-efficiency solar cell based on dye-sensitized colloidal TiO₂ films. *Nature* **353**, 737–740 (1991).
- Persson, P. *et al.* N 1s x-ray absorption study of the bonding interaction of bi-isonicotinic acid adsorbed on rutile TiO₂(110). *J. Chem. Phys.* **112**, 3945–3948 (2000).
- Chang, E. K., Rohlfing, M. & Louie, S. G. Excitons and optical properties of alpha-quartz. *Phys. Rev. Lett.* **85**, 2613–2616 (2000).
- Sandell, A. *et al.* Bonding of an isolated K atom to a surface: Experiment and theory. *Phys. Rev. Lett.* **78**, 4994–4997 (1997).
- Lu, H., Prieskorn, J. N. & Hupp, J. T. Fast interfacial electron transfer: Evidence for inverted region kinetic behavior. *J. Am. Chem. Soc.* **115**, 4927–4928 (1993).
- Ramakrishna, S. & Willig, F. Pump-probe spectroscopy of ultrafast electron injection from the excited state of an anchored chromophore to a semiconductor surface in UHV: A theoretical model. *J. Phys. Chem. B* **104**, 68–77 (2000).
- Petersson, Å., Ratner, M. & Karlsson, H. O. Injection time in the metal oxide-molecule interface calculated within the tight-binding model. *J. Phys. Chem. B* **104**, 8498–8502 (2000).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements

We are grateful for financial support to the Consortium on Clusters and Ultrafine Particles and to the ATOMICS Consortium, which are funded by Stiftelsen for Strategisk Forskning, to Göran Gustafssons Stiftelse, and to Vetenskapsrådet. We acknowledge the Swedish National Supercomputer Centre (NSC) for computer time, and the Group for Numerically-intensive Computations of the IBM Research Laboratory Zürich and the Abteilung Parrinello of MPI Stuttgart for help with the calculations.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.A.B. (e-mail: paul.bruhweiler@fysik.uu.se).

Ecosystem carbon loss with woody plant invasion of grasslands

Robert B. Jackson*, Jay L. Banner†, Esteban G. Jobbágy*, William T. Pockman*‡ & Diana H. Wall§

* Department of Biology and Nicholas School of the Environment and Earth Sciences, Duke University, Durham, North Carolina 27708-0340, USA

† Department of Geological Sciences, University of Texas at Austin, Austin, Texas 78712, USA

§ Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, Colorado 80523, USA

The invasion of woody vegetation into deserts, grasslands and savannas is generally thought to lead to an increase in the amount of carbon stored in those ecosystems. For this reason, shrub and forest expansion (for example, into grasslands) is also suggested to be a substantial, if uncertain, component of the terrestrial carbon sink^{1–14}. Here we investigate woody plant invasion along a precipitation gradient (200 to 1,100 mm yr^{−1}) by comparing carbon and nitrogen budgets and soil δ¹³C profiles between six pairs of adjacent grasslands, in which one of each pair was invaded by woody species 30 to 100 years ago. We found a clear negative relationship between precipitation and changes in soil organic carbon and nitrogen content when grasslands were invaded by woody vegetation, with drier sites gaining, and wetter sites losing, soil organic carbon. Losses of soil organic carbon at the wetter sites were substantial enough to offset increases in plant biomass carbon, suggesting that current land-based assessments may overestimate carbon sinks. Assessments relying on carbon stored from woody plant invasions to balance emissions may therefore be incorrect.

One-third to one-half of the Earth's land surface has been transformed by human action^{2,4}. Many continuing transformations exchange woody and herbaceous plants^{15–18}, including deforestation, desertification, and woody plant invasion (the expansion of woody species into grasslands and savannas). Shifting dominance among herbaceous and woody vegetation alters primary production, plant allocation, rooting depth and soil faunal communities, potentially metres underground^{15,18}, in turn affecting nutrient cycling and carbon storage^{19,20}. The two carbon pools most likely to change are woody plant biomass and soil organic matter, the dominant pool of carbon and nitrogen in grasslands^{10,20–23}. New woody biomass stores carbon in amounts that depend on the age, productivity and density of the stand. Changes in soil organic carbon (SOC) over the entire rooting zone are much harder to predict, and have the potential to enhance or offset biomass carbon gains, complicating projections of ecosystem carbon storage.

On the basis of a global analysis of more than 2,700 SOC profiles²⁰, we examined where increased biomass C in woody plants might be offset by SOC losses (see Methods). Across the global data set, the slope of the relationship between SOC and precipitation was 2.6 times higher for grassland vegetation than for shrublands/woodlands ($P = 0.001$; see Supplementary Information). Whereas grassland SOC was statistically indistinguishable from values for woody plants at 200 mm mean annual precipitation, woodlands had 43% less total SOC than grasslands at 1,000 mm ($P < 0.01$). Although suggestive, this analysis lacked a direct test of vegetation change independent of other covarying factors, including soil properties. So we also examined the direct effect of vegetation change at six paired grassland and invaded woody sites along a

‡ Present address: Department of Biology, University of New Mexico, Albuquerque, New Mexico 87131-1091, USA.

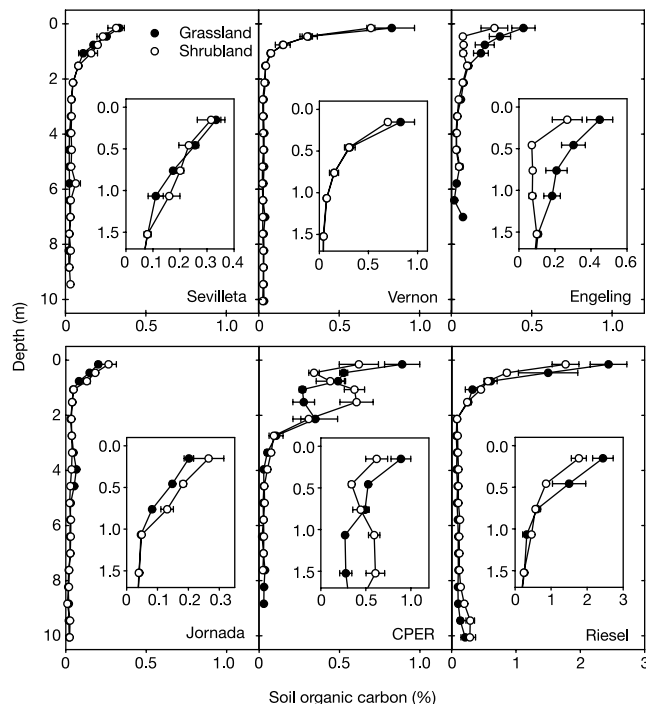


Figure 1 Concentrations of total soil organic carbon with depth. Data are shown for six paired grassland (filled circles) and woody sites (open circles) along the precipitation gradient (mean $\pm$ s.e.; $n = 3-8$). The panels are arranged from driest on the left to wettest on the right (analogous to the west-to-east precipitation gradient). The insets present the data for the top 1.5 m of soil at higher resolution. See Methods and Table 1 for additional information.

rainfall gradient in the southwestern USA (see Methods), testing relationships of biomass C gains and SOC losses for such common native woody invaders as *Prosopis* (mesquite), *Larrea* (creosote) and *Juniperus* (juniper) spp. (Table 1).

Along the precipitation gradient, woody plant invasion increased contents of SOC and soil organic nitrogen (SON) at the drier sites, and decreased them at the wetter sites (Fig. 1, Table 2). Jornada gained approximately one-quarter of total SOC after woody plant invasion (9 Mg ha^{-1}), but the two wettest sites lost one-fifth to almost one-half of total SOC (Table 2). Differential changes along the gradient resulted in a clear linear decline for the amounts of SOC and SON in the top 3 m of soil after conversion to woody plants (Fig. 2; $P \leq 0.01$ for both). Absolute changes in SOC and SON were also greater at the wetter sites; whereas none of the three desert sites gained more than 13 Mg ha^{-1} of SOC, the three wetter sites lost about 8, 32 and 61 Mg ha^{-1} (or about $0.25-0.8 \text{ Mg C ha}^{-1} \text{ yr}^{-1}$; Tables 1 and 2). The close agreement between the loss of SOC at Engeling (a 44% reduction in SOC at $1,070 \text{ mm}$ precipitation) and the database prediction (43% reduction in SOC at $1,000 \text{ mm}$

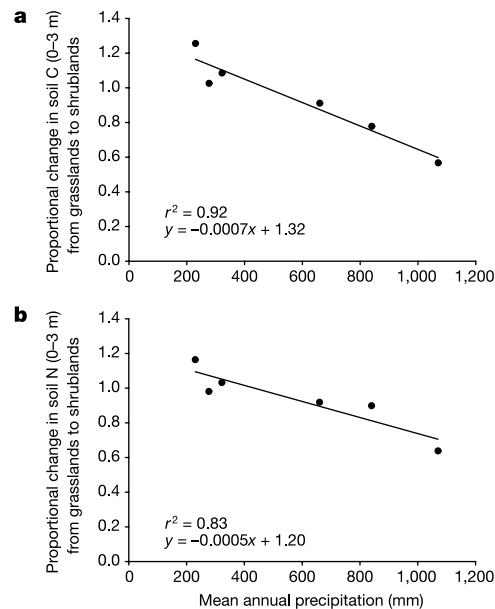


Figure 2 The proportional change in total soil organic carbon (a) and nitrogen (b) to 3 m depth with woody plant invasion of native grasslands. Values >1 are for those sites that gained soil organic carbon (SOC) and soil organic nitrogen (SON) after woody plant invasion, and values <1 denote a loss. See Table 1 for the key to the sites, mean annual precipitation, and the woody species.

precipitation) is notable.

SOC and SON were also more deeply distributed at the woody sites. Compared to the top metre of soil, there was 60% as much SOC at 1 to 3 m depth in the woodlands but only 40% in the grasslands (Table 2). Riesel lost 65 Mg C ha^{-1} and 3.9 Mg N ha^{-1} with woody plant invasion in the top metre of soil, but gained C and N slightly from 1 to 3 m depth; the CPER shrubland had 50% more SOC and SON between 1 and 3 m, but one-fifth less from 0 to 1 m compared to the grassland (Table 2).

Total ecosystem C stocks shifted both in size and in distribution above and below ground. Plant biomass C increased with woody plant invasion at all sites ($0.3-44 \text{ Mg C ha}^{-1}$), but the changes were smaller than for SOC at five of them (Table 2). The most notable shift in C from below-ground to above-ground pools was at the wettest site, Engeling, where 44 Mg C ha^{-1} of new plant C offset the SOC loss of 32 Mg ha^{-1} in the ~ 40 years since juniper invasion. Such shifts make C stocks more vulnerable to loss from fire, biomass harvesting and other disturbances. The biggest total change in C storage was for the Riesel tallgrass prairie site. There, a slight increase in plant biomass C was an order of magnitude smaller than SOC losses, resulting in a net loss of 56 Mg C ha^{-1} (Table 2). Total changes in ecosystem C pools at the other sites were more modest (-3 to 13 Mg C ha^{-1} ; Table 2).

Table 1 Descriptions and comparisons of the six study sites

Site	Location (°N, °W)	MAP* (mm)	Age† (yr)	Mean 95% rooting depth‡ (m)	Nematodes max. depth§ (m)	Mean depth of Sr uptake (m)
Jornada, NM	32.5, 106.8	230	>50	1.7, 5.4	1.8, 0.6	2.8, >3.0
Sevilleita, NM	34.3, 106.7	277	40	0.6, 2.0	0.6, 7.4	1.9, <1.2
CPER, CO	40.9, 104.7	322	>50	0.9, 4.0	2.5, 3.7	0.2, 1.4
Vernon, TX	33.9, 99.4	660	30	0.9, 5.8	4.3, 4.3	0.2, 0.1
Riesel, TX	31.5, 96.9	840	75-100	0.8, 2.8	1.8, 6.8	0.1, 2.4
Engeling, TX	31.9, 95.9	1,070	40	3.3, 2.5	1.2, 1.2	NA

* Mean annual precipitation.

† Approximate age of the woody stands at each site.

‡ Values given are for grasslands, woodlands.

§ Maximum depth of soil nematode occurrence at each site.

|| Integrated depth of soil Sr uptake ($^{87}\text{Sr}/^{86}\text{Sr}$ ratios; see Methods and Supplementary Information). NA, not available.

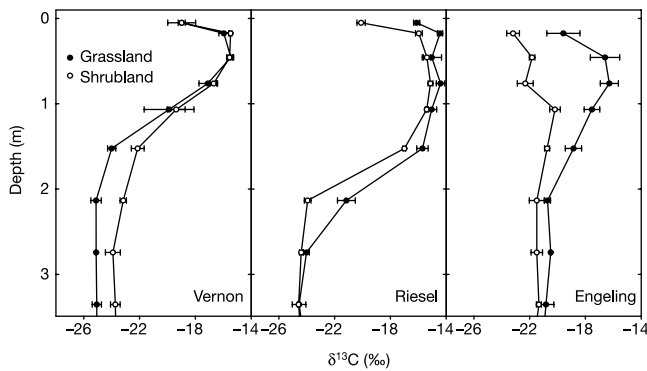


Figure 3 $\delta^{13}\text{C}$ values of soil organic carbon for the three wettest sites—Vernon, Riesel and Engeling. See Methods and Table 1 for additional information. Reference standard is Pee Dee Belemnite (V-PDB).

Examining in more detail the wetter sites where the greatest shifts occurred, we found that changes in C isotope signatures (Fig. 3) and N concentrations (see Supplementary Information) mirrored those for SOC pools (Fig. 1). The $\delta^{13}\text{C}$ data, reflecting differences in the photosynthetic pathway of C fixation for C_3 woody species and C_4 grasses¹⁷, show that the differences observed for grasslands and shrublands were not simply the result of losses of original SOC: new SOC was clearly incorporated into the system (Fig. 3). Using the assumptions of a two-layer mixing model (see Methods), ~57% of the SOC in the Engeling woodland came from the juniper trees in the ~40 years since invasion, with 43% of SOC in the woodland attributable to the original grassland SOC. At Riesel, about 20% of the SOC in the mesquite stand was attributable to the trees (Fig. 3).

Rooting depths, nematode distributions, and the integrated depth of soil cation uptake all changed substantially with woody plant invasion (Table 1). Mean 95% rooting depths were >2 m deeper on average for the woody vegetation at the six paired sites ($P \leq 0.01$), with increases of 3 to 4 m at Jornada, CPER and Vernon. Only Engeling showed no evidence of increased rooting depths with invasion. We also examined the representative soil biota most closely dependent on roots as indicator taxa—parasitic nematodes and nematodes feeding on rhizosphere bacteria and fungi. Their maximum depth of occurrence increased from 2.1 m on average in grasslands to 4.0 m for the woody sites (Table 1), and the composition of the nematode food web at this depth was markedly reduced from five trophic groups to two (bacterial- and root-feeding nematodes). Invaded woody sites also had lower species richness,

26 compared with 31 taxa, with the difference primarily due to the loss of root-feeding species. These responses of the nematode community to plant invasions are similar to changes in the soil food web with other disturbances, such as conversion to agriculture, that can alter rates of herbivory and decomposition²⁴.

We estimated the integrated depth of nutrient uptake using natural isotopes of Sr, an element geochemically similar to Ca (ref. 25) that is taken up by the same transporters in plants and undergoes negligible biological fractionation. Sr isotope ratios of the grasses at the wetter sites closely matched soil signatures in the shallowest layers, but at the two driest grasslands the integrated depth of uptake was a surprising 2 to 3 m (Table 1; see Supplementary Information). Just as importantly, woody plant invasion increased the depth of Sr uptake at three sites by as much as 2 m (Table 1).

Our global database analysis and fieldwork both suggest an important role for precipitation in understanding the consequences of desertification and woody plant invasions. Nonetheless, the database also showed significant unexplained variation in SOC storage^{13,20,26,27}. Soil factors can influence SOC concentrations and the occurrence of woody encroachment²⁸. In a south Texas savanna, for example, soil texture and the presence of an argillic horizon strongly influence where *Prosopis* establishes²⁸. The condition and productivity of grasslands before invasion are also important. Relatively wet grasslands are often highly productive, allocating a large proportion of C below ground, and have high SOC concentrations^{13,20,23}. An analysis of 115 studies concluded that sites originally dominated by woodlands showed the greatest potential of all biomes for storing C as managed pastures²⁹, in part because of high grassland productivity and below-ground allocation there. A similar meta-analysis of land conversion from managed pastures to plantations in New Zealand also documented soil carbon losses¹³. Historical factors and disturbance, including erosion and fire frequency, play additional roles^{26,27}. Modelling simulations of the *Prosopis* savanna mentioned above²⁸ suggest that one-fifth of SOC could be lost with grazing before woody plant invasion. Of these many potential factors, our experimental design isolated the role of vegetation change for C storage and ecosystem functioning in the absence of background differences in soils.

Potential C storage in plant biomass is common as woody plants invade grasslands. Nonetheless, current uncertainties for net changes in C stocks are large, as are the uncertainties in regional extrapolations of woody plant invasions^{11,12}. Recent analyses have nevertheless concluded that the US carbon budget is approximately balanced¹¹. Our data suggest that this conclusion may be premature.

Table 2 Soil organic C and N to 3 m depth and total changes in C pools after woody plant invasion

Site	Depth	Soil organic carbon (Mg ha^{-1})			Soil organic nitrogen (Mg ha^{-1})			Net Δ C (Mg ha^{-1})	
		Grassland	Shrub- or woodland	Change (%)	Grassland	Shrub- or woodland	Change (%)		
Jornada	0–1 m	24.6	32.9	+33%	3.2	3.8	+21%	Soil	8.7
	1–3 m	9.2	9.6	+5%	1.3	1.3	+6%	Plant	1.4
	0–3 m*	33.8	42.5	+26%	4.4	5.2	+16%	Total	10.1
Sevilleta	0–1 m	38.0	38.0	0%	4.8	4.7	–4%	Soil	1.4
	1–3 m	14.6	16.0	+10%	1.5	1.5	+4%	Plant	0.4
	0–3 m*	52.6	54.0	+3%	6.3	6.2	–2%	Total	1.8
CPER	0–1 m	89.3	70.8	–21%	10.5	7.6	–27%	Soil	12.8
	1–3 m	60.8	92.1	+52%	6.2	9.6	+55%	Plant	0.3
	0–3 m*	150	163	+9%	16.7	17.3	+3%	Total	13.1
Vernon	0–1 m	72.4	65.0	–10%	9.4	8.4	–11%	Soil	–7.6
	1–3 m	13.7	13.5	–1%	5.4	5.2	–4%	Plant	4.7
	0–3 m*	86.1	78.5	–9%	14.8	13.6	–8%	Total	–2.9
Riesel	0–1 m	229	164	–28%	20.2	16.4	–19%	Soil	–61.4
	1–3 m	48.4	51.9	+7%	6.2	7.4	+19%	Plant	5.3
	0–3 m*	278	216	–22%	26.4	23.7	–10%	Total	–56.1
Engeling	0–1 m	47.9	20.7	–57%	4.0	1.6	–59%	Soil	–32.2
	1–3 m	26.0	21.0	–20%	2.9	2.8	–5%	Plant	44.4
	0–3 m*	73.9	41.7	–44%	6.9	4.4	–37%	Total	12.2

* Values of SOC and SON for 0–3 m depth are the sum of the respective values at 0–1 and 1–3 m depth.

On the basis of the database analysis presented here and on the observed losses of SOC in the wetter grasslands, further field studies and modelling integration will be needed to evaluate the net effects of woody plant invasions for the C cycle and to close the C budget. □

Methods

Database analysis

The global analysis of the variation of SOC with precipitation was derived from the National Soil Characterization Database of the USDA and the World Inventory of Soil Emission Potential Database of the International Soil Reference and Information Centre (see ref. 20). Mean annual precipitation for each location was obtained with a resolution of $0.5^\circ \times 0.5^\circ$ from the International Institute for Applied Systems Analysis. The analysis considered soils to 1-m depth (insufficient data existed for deeper depths) at all sites with <1,000 mm mean annual precipitation except Histosols (for example, organic soils in bogs and marshes) and Udi/Ustipsamments (dune soils). Herbaceous-dominated communities excluding riparian meadows and tundra were considered grasslands, and woodland, brush, chaparral, scrub, shrub steppe, or shrub desert communities were considered woody sites. The final regression statistics for SOC (kg C m^{-2}) with mean annual precipitation (mm) were as follows. Grasslands: y intercept, 3.253; slope, 0.0152; $r^2 = 0.33$; $P < 0.001$; $n = 75$. Woody vegetation: y intercept, 4.629; slope, 0.00595; $r^2 = 0.12$; $P < 0.001$; $n = 167$ (see Supplementary Information). Because many factors change or are correlated with rainfall, this analysis merely suggested a strong interaction with precipitation.

Experimental design and soil sampling

Field measurements were made in adjacent communities at the six locations in Table 1. The 200,000-ha Waggoner ranch is the long-term research site for the Vernon, Texas, Research Station of Texas A&M University. The Riesel site maintained by the USDA/Agriculture Research Service, and the Gus Engeling Wildlife Management Area operated by the Texas Parks and Wildlife Department, have been maintained by these agencies for >50 years. The other sites are part of the international Long Term Ecological Research (LTER) network. On the basis of aerial photographs and additional records, the youngest age of the woody stands was 30 to 40 years (Sevillea, Vernon and Engeling). The dominant grasses at each site were *Bouteloua eriopoda* at Jornada and Sevillea, *Bouteloua gracilis* at CPER, *Stipa* sp. at Vernon, *Schizachyrium scoparium* at Riesel, and *Andropogon* sp. at Engeling.

Each site contained adjacent grassland and shrub/woodland communities where management practices have contributed to woody plant expansion into native grasslands on one side of a fence. The goal of this paired experimental design was to compare the effects of herbaceous and woody communities without initial differences in soil properties. The one site where initial differences may have existed was the Shortgrass Steppe LTER (CPER), as revealed by the background soil profiles of Sr isotopes (see Supplementary Information).

In each of the two communities per site, four to nine 6-cm-diameter cores were taken with an environmental drilling rig to 10-m depth. The cores were extracted between May and October 1997 (the approximate time of peak biomass at each site) in 61-cm increments, and were subsequently air-dried and passed through a 2-mm sieve, where roots were extracted and quantified. To remove soil inorganic C (present in a subset of samples at all sites except Engeling), soil samples were treated with 1 N $\text{H}_2\text{SO}_4/5\%$ FeSO_4 and double-checked for complete inorganic C removal. Total SOC and N were measured as in ref. 30 using a CE Instruments NC 2100 elemental analyser (ThermoQuest Italia), with a subset of measurements confirmed independently. The $\delta^{13}\text{C}$ SOC measurements were run on a Finnigan MAT DeltaPlusXL mass spectrometer at Duke University. The grasslands generally had a mixture of C_3 and C_4 species, so the grassland $\delta^{13}\text{C}$ signatures were used as one end of a two-layer mixing model for the approximate calculations of C derived from the woody plants after invasion. The shrubs were estimated to have a $\delta^{13}\text{C}$ signature of -26% , typical for C_3 plants. Total SOC and N concentrations were converted to ecosystem estimates using the bulk density at each depth and correcting for the presence of any rocks (>2 mm) (see Supplementary Information). To help address any spatial heterogeneity and to supplement data from the deep cores, 25-cm-diameter soil pits were dug to 50-cm depth randomly in each community ($n \geq 4$ per community). The soil subsamples for nematodes were sent immediately to Colorado State University for analysis²⁴. Sr isotope analyses of soil and plant material were determined at the University of Texas at Austin using a Finnigan-MAT 261 thermal ionization mass spectrometer in automated, dynamic multi-collection mode (see Supplementary Information).

Proportional changes in total SOC and SON to 3-m depth with woody plant invasion (Fig. 2) were analysed by regression ($n = 6$), using precipitation as the dependent variable. Other statistical determinations were calculated by paired t -test.

Plant biomass C

Herbaceous biomass in grassland communities was estimated by harvesting all material in random $0.5 \text{ m} \times 0.5 \text{ m}$ plots ($n \geq 4$ per community). Live and dead standing biomass were separated, oven dried, weighed, and converted to biomass C. At the drier end of the precipitation gradient, significant herbaceous biomass also occurred in the shrubland communities. There, four under-canopy and four open positions were randomly selected for $0.5 \text{ m} \times 0.5 \text{ m}$ herbaceous harvest plots. The proportions of under-canopy and open positions were estimated using three randomly located 60-m Canfield transects per community, with mean herbaceous biomass adjusted by the proportional cover of each position. For woody plant biomass, the density of plants was determined in each of three

random $10 \text{ m} \times 20 \text{ m}$ plots per community, with plant biomass harvested in each plot. Wood and leaf samples were separated, oven dried, weighed, and converted from total biomass to biomass C. Fine root biomass estimates were obtained from the soil cores after sieving, drying and ashing the root samples. Large woody roots (>1 cm diameter) were the only C pool not quantified in this study.

Received 14 January; accepted 14 June 2002; doi:10.1038/nature00910.

- Schlesinger, W. H. *et al.* Biological feedbacks in global desertification. *Science* **247**, 1043–1048 (1990).
- Vitousek, P. M., Mooney, H. A., Lubchenco, J. & Melillo, J. M. Human domination of Earth's ecosystems. *Science* **277**, 494–499 (1997).
- Schimel, D. S. *et al.* Recent patterns and mechanisms of carbon exchange by terrestrial ecosystems. *Nature* **414**, 169–172 (2001).
- Turner, B. L. II *et al.* (eds) in *The Earth as Transformed by Human Action* (Cambridge Univ. Press, Cambridge, 1990).
- Tucker, C. J., Dregne, H. E. & Newcomb, W. W. Expansion and contraction of the Sahara Desert from 1980 to 1990. *Science* **253**, 299–301 (1991).
- Scholes, R. J. & Archer, S. R. Tree-grass interactions in savannas. *Annu. Rev. Ecol. Syst.* **28**, 517–544 (1997).
- Brown, J. H., Valone, T. J. & Curtin, C. G. Reorganization of an arid ecosystem in response to recent climate change. *Proc. Natl Acad. Sci. USA* **94**, 9729–9733 (1997).
- Allen, C. D. & Breshears, D. D. Drought-induced shift of a forest-woodland ecotone: rapid landscape response to climate variation. *Proc. Natl Acad. Sci. USA* **95**, 14839–14842 (1998).
- Casper, B. B. & Jackson, R. B. Plant competition underground. *Annu. Rev. Ecol. Syst.* **28**, 545–570 (1997).
- Amundson, R. The carbon budget in soils. *Annu. Rev. Earth Planet. Sci.* **29**, 535–562 (2001).
- Pacala, S. W. *et al.* Consistent land- and atmosphere-based U.S. carbon sink estimates. *Science* **292**, 2316–2320 (2001).
- Houghton, R. A., Hackler, J. L. & Lawrence, K. T. The U. S. carbon budget: contributions from land-use change. *Science* **285**, 574–578 (1999).
- Guo, L. B. & Gifford, R. M. Soil carbon stocks and land use change: a meta analysis. *Glob. Change Biol.* **8**, 345–360 (2002).
- Fang, J. Y., Chen, A. P., Peng, C. H., Zhao, S. Q. & Longjun, C. Changes in forest biomass carbon storage in China between 1949 and 1998. *Science* **292**, 2320–2322 (2001).
- Nepstad, D. C. *et al.* The role of deep roots in the hydrological and carbon cycles of Amazonian forests and pastures. *Nature* **372**, 666–669 (1994).
- Van Auken, O. W. Shrub invasions of North American semiarid grasslands. *Annu. Rev. Ecol. Syst.* **31**, 197–215 (2000).
- Boutton, T. W. *et al.* Delta C-13 values of soil organic carbon and their use in documenting vegetation change in a subtropical savanna ecosystem. *Geoderma* **82**, 5–41 (1998).
- Jackson, R. B. *et al.* Belowground consequences of vegetation change and their treatment in models. *Ecol. Appl.* **10**, 470–483 (2000).
- Trumbore, S. E. Potential responses of soil organic carbon to global environmental change. *Proc. Natl Acad. Sci. USA* **94**, 8284–8291 (1997).
- Jobbágy, E. G. & Jackson, R. B. The vertical distribution of soil organic carbon and its relation to climate and vegetation. *Ecol. Appl.* **10**, 423–436 (2000).
- Post, W. M., Emanuel, W. R., Zinke, P. J. & Stangenberger, A. G. Soil carbon pools and world life zones. *Nature* **298**, 156–159 (1982).
- Batjes, N. H. Total carbon and nitrogen in the soils of the world. *Eur. J. Soil Sci.* **47**, 151–163 (1996).
- Neill, C. & Davidson, E. A. in *Global Climate Change and Tropical Ecosystems* (eds Lal, R., Kimble, J. M. & Stewart, B. A.) 197–211 (CRC, Boca Raton, 2000).
- Wall, D. H., Adams, G. A. & Parsons, A. N. in *Global Biodiversity in a Changing Environment* (eds Chapin, F. S., Sala, O. E. & Huber-Sannwald, E.) 47–82 (Springer, New York, 2001).
- Vitousek, P. M., Kennedy, M. J., Derry, L. A. & Chadwick, O. A. Weathering versus atmospheric sources of strontium in ecosystems on young volcanic soils. *Oecologia* **121**, 255–259 (1999).
- Burke, I. C. *et al.* Texture, climate, and cultivation effects on soil organic matter content in U.S. grassland soils. *Soil Sci. Soc. Am. J.* **53**, 800–805 (1989).
- Tilman, D. *et al.* Fire suppression and ecosystem carbon storage. *Ecology* **81**, 2680–2685 (2000).
- Archer, S., Boutton, T. W. & Hibbard, K. A. in *Global Biogeochemical Cycles in the Climate System* (eds Schulze, E.-D. *et al.*) 115–137 (Academic, San Diego, 2001).
- Conant, R. T., Paustian, K. & Elliott, E. T. Grassland management and conversion to grassland: effects on soil carbon. *Ecol. Appl.* **11**, 343–355 (2001).
- Gill, R. A. *et al.* Nonlinear grassland responses to past and future atmospheric CO_2 . *Nature* **417**, 279–282 (2002).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements

We thank W. Cook, L. Giles, J. Karr, L. Mack, A. Parsons and S. Rainey for laboratory analyses, and W. H. Schlesinger, A. T. Austin, O. E. Sala and E. A. Davidson for manuscript suggestions. We also thank R. J. Ansley, H. W. Polley and many others who helped us locate sites and provide their history. This work was supported by the US National Science Foundation, NIGEC/DOE, the Inter-American Institute for Global Change Research, the Andrew W. Mellon Foundation, and the Geology Foundation of the University of Texas at Austin. This paper is a contribution to the Global Change and Terrestrial Ecosystems core project of the International Geosphere Biosphere Programme.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.B.J. (e-mail: jackson@duke.edu).

Reassessing the evidence for the earliest traces of life

Mark A. van Zuilen*, Aivo Lepland†‡ & Gustaf Arrhenius*

* Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California 92093-0236, USA

† Geological Survey of Norway, Leiv Eirikssons vei 39, 7491 Trondheim, Norway

‡ Institute of Geology, Tallinn Technical University, EE-0001 Tallinn, Estonia

The isotopic composition of graphite is commonly used as a biomarker in the oldest (>3.5 Gyr ago) highly metamorphosed terrestrial rocks. Earlier studies on isotopic characteristics of graphite occurring in rocks of the approximately 3.8-Gyr-old Isua supracrustal belt (ISB) in southern West Greenland have suggested the presence of a vast microbial ecosystem in the early Archean^{1–4}. This interpretation, however, has to be approached with extreme care⁵. Here we show that graphite occurs abundantly in secondary carbonate veins in the ISB that are formed at depth in the crust by injection of hot fluids reacting with older crustal rocks (metasomatism). During these reactions, graphite forms from the disproportionation of Fe(II)-bearing carbonates at high temperature. These metasomatic rocks, which clearly lack biological relevance, were earlier thought to be of sedimentary origin and their graphite association provided the basis for inferences about early life^{1–4}. The new observations thus call for a reassessment of previously presented evidence for ancient traces of life in the highly metamorphosed Early Archean rock record.

Metasedimentary units, including detrital conglomerates, turbidites, chemically precipitated banded iron formations (BIFs) and metacherts, as well as pillow structures of mafic rocks, indicate the accumulation of ISB rocks in an aquatic environment^{6–10}. Carbonate-rich Isua rocks have also originally been interpreted as being part of the sedimentary succession^{10,11}. However, recent studies have disproved this interpretation^{6,7,9,12–14}. Metacarbonate deposits and calc-silicate rocks bear signs of having formed by metasomatic reactions during fluid infiltration across contacts between igneous ultramafic rocks and their host rocks^{12–14}. It is now widely accepted that most, if not all, carbonate in Isua is metasomatic and not sedimentary in origin^{6,7,9}.

The light isotopic signatures ($\delta^{13}\text{C} \approx -25\text{‰}$) of reduced carbon in some of the Isua rocks have been used as evidence for biologic activity in the early Archean^{2–4}. Apart from these few observed light carbon isotope signatures, genetically less indicative bulk carbon isotope ratios (range from -25 to -6‰) were also obtained, particularly from carbonate-rich ISB rocks^{2,15–17}. Ion microprobe analysis of individual graphite particles has yielded $\delta^{13}\text{C}$ values from -18 to $+2\text{‰}$ and has shown isotopically heterogeneous graphite within a single rock specimen¹⁸.

Metamorphically induced isotope exchange between carbonate and graphite in these rocks was thought to have caused the partial loss of a conjectured biogenic light carbon isotope signature of the kind that is typically found in younger, less metamorphosed organic rich sediments in the geological record^{2–4}. Furthermore, it has been shown that the original light isotope signature of organic matter is gradually diminished by volatilization during diagenetic and metamorphic alteration to kerogen and ultimately lost on conversion to graphite¹⁵. No kerogen but only crystalline graphite is found in Isua rocks, complicating the possible biogenic interpretation of graphite in these rocks¹⁵.

Micrometre-size graphite inclusions with a pronounced light $\delta^{13}\text{C}$ value (weighted mean $-30 \pm 3\text{‰}$; ion microprobe data) were reported to occur in apatite crystals in one sample from the ISB¹. The graphite was thought to have escaped isotope exchange with the

associating carbonates by means of armouring by the host apatite¹, which seems to strengthen the evidence of early life in these rocks. The apatite itself was considered as a potential biomarker¹, in analogy with phosphate precipitates found in organic rich sedimentary deposits¹⁹ such as phosphorites.

The recognition of the secondary, metasomatic origin of carbonate rich rocks in the ISB has highlighted the need for assessment of graphite genesis and its applicability to the interpretation of life at 3.8 Gyr ago. We have for this purpose studied a suite of samples from the ISB including metacarbonates, turbidites, cherts and BIFs (see Methods section for details).

Scanning electron microscope-energy dispersive spectrometer (SEM-EDS) analysis of carbonate phases in the Isua rocks reveals three distinct carbonate cation compositions (Fig. 1). Iron-rich carbonate (MgMn-siderite) was only found in metacarbonate samples from veins occurring as distinct layers in mafic metavolcanic country rock. Compositions with less iron (Fe-dolomite) were found in one metacarbonate layer occurring within metachert (AL17), and in small veinlets within some BIFs and metacherts. Iron-deficient carbonate (calcite) was found in veinlets within BIF, metachert and turbidite sequence samples. These different compositions are likely to reflect the interaction between the country rock and fluid, and explain the dominance of MgMn-siderite in metacarbonate veins within the iron-rich metavolcanic rocks. In field appearance and mineralogy these veins closely resemble magnetite bearing, quartz-poor, highly graphitic bands that were previously considered in ref. 20 to be a carbonate facies BIF, but are now recognized as metasomatic carbonates.

Graphite is relatively abundant (up to approximately 2 wt%) in metacarbonate samples (Fig. 2a), whereas the concentration of reduced carbon in our metasedimentary BIF and metachert samples is below 100 p.p.m.. Micropetrographic analysis reveals that the graphite in metacarbonates occurs closely associated with Fe-bearing carbonate and magnetite or FeMg-amphibole (silicate). The association of graphite with MgMn-siderite and magnetite (Fig. 3a) was observed in all graphitic veins within mafic country rocks, suggesting that graphite and magnetite in these rocks are the

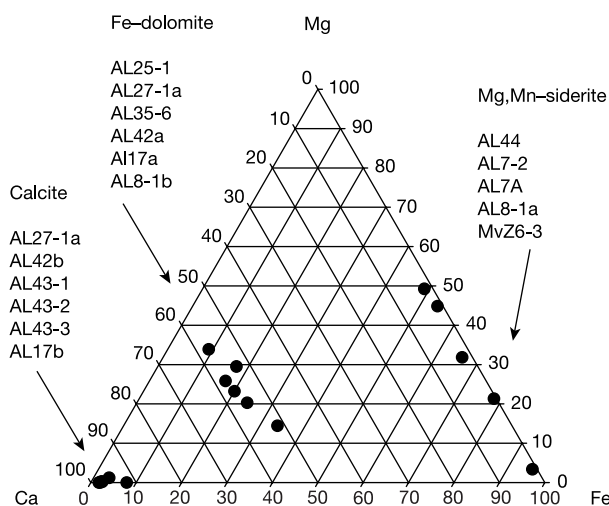


Figure 1 Carbonate phases can be divided into three distinct populations. These are: calcite ($\text{Fe}_{0.02-0.08}\text{Mg}_{0.00-0.03}\text{Mn}_{0.01-0.04}\text{Ca}_{0.89-0.97}\text{CO}_3$), Fe-dolomite ($\text{Fe}_{0.09-0.36}\text{Mg}_{0.14-0.34}\text{Mn}_{0.00-0.06}\text{Ca}_{0.50-0.57}\text{CO}_3$), and MgMn-siderite ($\text{Fe}_{0.40-0.85}\text{Mg}_{0.03-0.40}\text{Mn}_{0.09-0.22}\text{CO}_3$). AL8-1, AL17, AL27-1 and AL42 contain carbonate minerals with two different compositions (populations a and b) and support an earlier suggestion of the occurrence of multiple pulses of carbonate metasomatism in the ISB¹². Not shown are AL13-1, AL15-1B, MvZ6.4D1, MvZ6.4D2, because they do not contain carbonate. Samples MvZ6.4D1 and MvZ6.4D2 are highly graphitic, chlorite and magnetite dominated interlayers within MgMn-siderite-rich metacarbonate veins.

products of thermal disproportionation of the carbonate. Iron is present in reduced (divalent) form in these carbonates, and is oxidized to form magnetite, and the carbonate ion is reduced to graphite. Similar Fe-carbonate dissociation reactions apparently yield graphite in veins within Si-rich metacherts (sample AL17), but the reaction product in the presence of quartz is Fe-silicate rather than magnetite (compare ref. 8). The siderite (FeCO_3) disproportionation reaction, yielding graphite and magnetite has been studied in detail²¹ at metamorphic $P,T,f\text{O}_2$ conditions, and has been suggested earlier as a possible mechanism for graphite formation in the amphibolite facies ($T \approx 550^\circ\text{C}$; $P \approx 5$ kbar) ISB¹⁷. Thermal decomposition of pure siderite occurs at temperatures above 450°C in the reaction:



The isotopic composition of bulk graphite in the metacarbonates ranges from -12 to -10‰ (Fig. 2b), which is significantly heavier than graphitized kerogen (typically near $\delta^{13}\text{C} \approx -25\text{‰}$ in Early Archaean rocks¹⁵). The isotopic composition of the carbonate carbon ranges between $\delta^{13}\text{C}$ values of $+1$ and -9‰ (ref. 22). Most of the graphite and associated Fe-carbonate in metacarbonate samples display a carbon isotopic difference ($\Delta_{\text{graphite-carbonate}}$) close to -6‰ , in accord with the calculated isotope equilibrium fractionation between the two phases at around $1,500^\circ\text{C}$ (ref. 23) (assuming $\Delta_{\text{graphite-siderite}}$ to be equal to $\Delta_{\text{graphite-calcite}}$). This matches the estimated temperature range of metamorphism that

these rocks have experienced²⁴. We therefore conclude that graphite in these rocks obtained a $\delta^{13}\text{C}$ signature between -12 and -10‰ owing to equilibrium isotope fractionation during the thermal dissociation of metasomatic Fe-carbonates. The occurrence of graphitic metacarbonates as distinct veins, the association of graphite with Fe-bearing carbonates and magnetite or FeMg-amphibole, and its isotopic composition of $\delta^{13}\text{C} \approx -12\text{‰}$ are all consistent with an abiogenic origin of graphite by carbonate reduction.

The earlier study² on the isotopic composition of graphite in Isua was focused on carbonate-rich rocks that at the time were believed to have a sedimentary origin. The range of isotopic compositions of reduced carbon ($\delta^{13}\text{C}$ from -6 to -25‰) found in ref. 2 is inconsistent with our data that show $\delta^{13}\text{C}$ values of graphite in a narrow range between -12 and -10‰ in these rocks. With the current view of the metasomatic origin of carbonate-rich rocks^{12–14} and the mechanism of graphite formation in such rocks as shown in this study, the biogenic interpretation given in ref. 2 seems to be open to reassessment.

No distinguishable graphite particles were found in our sedimentary BIF and metachert samples and combustion measurements indicated extremely low (less than 100 p.p.m.) concentrations of reduced carbon. We measured the $\delta^{13}\text{C}$ of reduced carbon in these samples and found $\delta^{13}\text{C}$ values consistently between -25 and -30‰ (Fig. 2b). In all of these samples the major fraction of reduced carbon combusted at a 450°C step (shown for AL25 in Fig. 4). Graphite typically combusts at around $700\text{--}800^\circ\text{C}$, so we conclude that the small amount of isotopically light, reduced carbon is mainly derived from unmetamorphosed recent organic material, which survives acid treatment and cleaning in organic solvents during sample processing. This organic matter is possibly

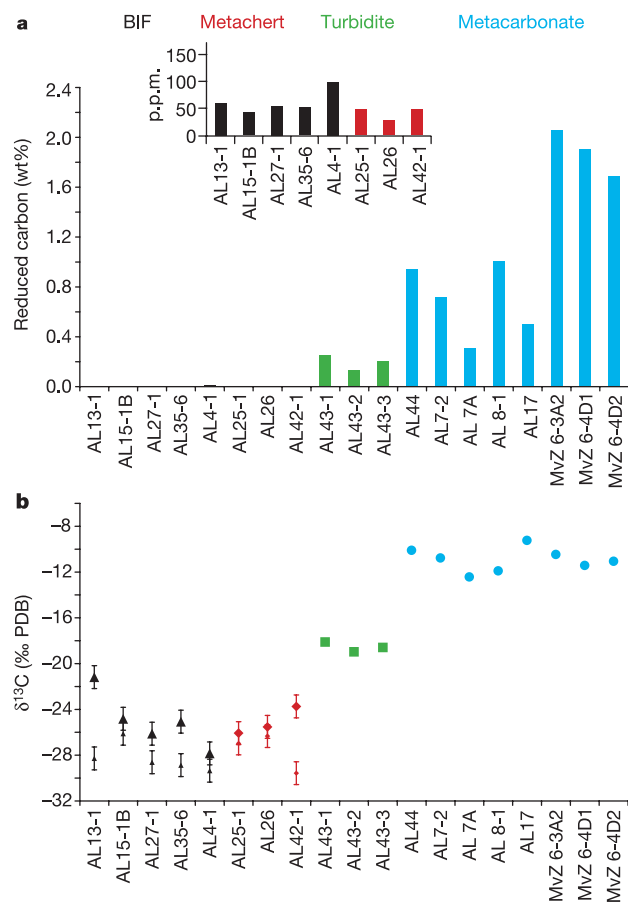


Figure 2 Abundance and isotope composition of reduced carbon. **a**, Abundances (wt%) of reduced carbon in studied samples. In a separate plot a close-up of low concentrations (carbon content in p.p.m.) in BIFs and metacherts is shown. **b**, Carbon isotope composition of total reduced carbon in BIFs (black triangles), metacherts (red diamonds), turbidites (green squares), metacarbonates (blue circles). Isotopic composition for carbon released at 450°C is shown for BIFs (small black triangles) and metacherts (small red diamonds).

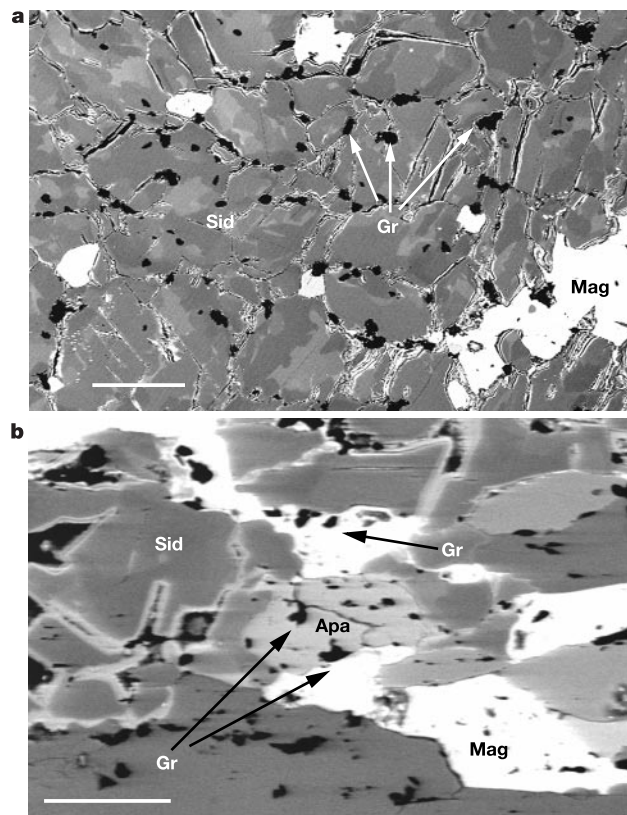


Figure 3 Backscattered electron (BSE) images of samples. **a**, BSE image of AL44, showing the spatial distribution of graphite (Gr) in relation to MgMn-siderite (Sid) and magnetite (Mag). Scale bar, $200\text{ }\mu\text{m}$. **b**, BSE image of I-3381, showing an apatite crystal with graphite inclusions (Apa) in relation to magnetite, MgMn-siderite and graphite. Scale bar, $50\text{ }\mu\text{m}$.

emplaced by percolating ground water²⁵. Because almost no carbon phase is found to combust at temperatures above 450 °C (Fig. 4), we conclude that the typical $\delta^{13}\text{C} \approx -28\text{‰}$ signature in these graphite-deficient sedimentary rocks is entirely due to recent contamination. We therefore question the palaeobiological significance of earlier single-step combustion experiments on decarbonated bulk rock samples from Isua.

The ion microprobe data on graphite in Isua metacherts and psammitic schists¹⁸ revealed particles with $\delta^{13}\text{C}$ values as light as -18‰ . These light isotopic compositions cannot be related to recent contamination; they may indeed constitute ancient traces of life. However, the wide range of $\delta^{13}\text{C}$ values and common occurrence of graphite in samples containing carbonate and products of carbonate replacement (amphiboles) calls for further petrographic and isotopic studies and critical evaluation of abiogenic decarbonation processes before a biogenic interpretation of the origin of graphite in these samples¹⁸ can safely be made.

The biogenic interpretation of isotopically light graphite inclusions in apatite crystals in the ISB was based on a single rock sample I-3381 (ref. 1), believed at the time to represent a sedimentary BIF. The recent petrographic analysis using the original thin section of I-3381 has revealed that it contains MgMn-siderite-magnetite-graphite associations and is compositionally akin to Isua metacarbonates²⁶. The compositional similarity of I-3381 and AL8-1, as well as the field descriptions, indicate that these two samples were collected from the same metacarbonate outcrop²⁶. In all metacarbonate samples, including I-3381, graphite is not restricted to apatite, but occurs as inclusions in most other phases too (Fig. 3b). The petrographic evidence strongly suggests that the origin of graphite in apatite is no different from graphite included in other phases and, regardless of mineral association, is produced epigenetically through thermal disproportionation of ferrous carbonate. This type of graphite consequently appears to be without biological significance and formed in one or several thermal events later than 3.8 Gyr ago. The isotopic systematics of the process responsible for formation of isotopically light graphite (weighted mean $-30 \pm 3\text{‰}$) enclosed in apatite crystals in I-3381 (ref. 1) remains to be studied, but our petrographic evidence clearly excludes a primary biogenic origin.

A sample from granulite facies rocks from Akilia island, around 150 km southwest of Isua, has also been reported to contain isotopically light graphite inclusions in apatite crystals¹. The present study does not include Akilia rocks and our conclusions about the origin of graphite and apatite-graphite association in Isua cannot, without additional petrographic and chemical analyses, be applied to the more extensively metamorphosed rocks from Akilia island, whose protolith and age are currently debated²⁷.

Significant amounts of graphite (>0.1 wt%) were found in rocks

interpreted as graded beds (turbidite), in the western part of the belt²⁸ (Fig. 2a). The isotopic composition of graphite in three samples from the turbidite sequence is in the present work found to be between -18 and -19‰ (Fig. 2b), consistent with the original measurements²⁸. Graphite in these samples does not coexist with siderite or magnetite; in fact no iron-containing carbonate phases were found in these rocks (Fig. 1), indicating that siderite disproportionation is an unlikely mechanism for graphite production. The original biogenic interpretation of graphite in this formation is therefore not challenged by our observations.

Our study leads to conclusions that require basic modification of some earlier inferences about traces of life in the ancient Isua rocks. A biogenic origin of graphite in carbonate-rich rocks in Isua¹⁻⁴ was inferred from the assumption that these rocks had a sedimentary origin. However, recent field and laboratory investigations have shown that most if not all carbonate in Isua is metasomatic in origin. Petrographic and isotopic analyses show that graphite in the metacarbonate rocks, serving as a basis for earlier investigations, is produced abiogenically by disproportionation of ferrous carbonate at high temperature and pressure and at a time later than the formation of the host rock. This type of graphite, including graphite inclusions in apatite, therefore cannot represent 3.8 Gyr-old traces of life. Stepped-temperature combustion accompanied by isotope analysis shows that single-step combustion of bulk rock samples cannot be used for tracing life in graphite-deficient metasediments. □

Methods

Rock samples collected from a zone of relatively low strain⁶ in the northeastern part of the ISB include metacarbonates (AL17, AL7-2, AL8-1, AL7A, MvZ6.3A2, MvZ6.4D1, MvZ6.4D2), BIFs (AL4-1, AL27-1, AL13, AL35-6, AL15) and metacherts (AL25, AL26, AL42). From the western part of the belt a metacarbonate sample (AL44) was collected, as well as three samples from a metasedimentary turbidite sequence²⁸ (AL43-1, AL43-2, AL43-3). We apply the term 'metacarbonate' to cm to m scale carbonate-rich veins, typically cross-cutting the country rock. The metasedimentary samples, described as BIFs and metacherts, also contain small amounts ($<5\%$) of carbonate minerals in small veinlets, but the texture and composition of these rocks have not changed significantly as a result of metasomatism.

Analysis of the cation composition of the carbonate minerals was performed with an EDS in a Cambridge 360 SEM using polished thin sections. Samples were prepared for carbon isotope analysis by decarbonation (4 M HCl solution at 70 °C for 48 h) and cleaning in a 10:1 dichloromethane/ethanol solution. Carbon concentration was determined with a Perkin-Elmer 2400 CHN elemental analyser (graphitic samples), or determined by capacitance manometry from CO_2 release during stepped-heating combustion (graphite-deficient samples). Conversion to CO_2 was achieved by sealed tube combustion (graphitic samples) or by stepped-heating combustion (graphite-deficient samples). After combustion, CO_2 was collected cryogenically (procedure described in ref. 29) and carbon isotopes were measured on a VG PRISM multicollector mass spectrometer. All measurements are Craig-corrected and presented as $\delta^{13}\text{C}$ values ($\delta^{13}\text{C} = ((^{13}\text{C}/^{12}\text{C})_{\text{sample}} / (^{13}\text{C}/^{12}\text{C})_{\text{standard}}) - 1) \times 1,000$) relative to Pee-Dee Belemnite (PDB). USGS24 and NBS19 were used as standards.

Received 16 April; accepted 11 June 2002; doi:10.1038/nature00934.

1. Mojzsis, S. J., Arrhenius, G. *et al.* Evidence for life on Earth before 3,800 million years ago. *Nature* **384**, 55–59 (1996).
2. Schidlowski, M., Appel, P. W. U., Eichmann, R. & Junge, C. E. Carbon isotope geochemistry of the 3.7–10⁹-yr-old Isua sediments, West Greenland: implications for the Archaean carbon and oxygen cycles. *Geochim. Cosmochim. Acta* **43**, 189–199 (1979).
3. Schidlowski, M. A 3,800-million-year isotopic record of life from carbon in sedimentary rocks. *Nature* **333**, 313–318 (1988).
4. Schidlowski, M. Carbon isotopes as biogeochemical recorders of life over 3.8 Ga of Earth history: evolution of a concept. *Precamb. Res.* **106**, 117–134 (2001).
5. Hayes, J. M. The earliest memories of life on Earth. *Nature* **384**, 21–22 (1996).
6. Appel, P. W. U., Fedo, C. M., Moorbath, S. & Myers, J. S. Recognizable primary volcanic and sedimentary features in a low-strain domain of the highly deformed, oldest known (~3.7–3.8 Gyr) Greenstone Belt, Isua, West Greenland. *Terra Nova* **10**, 57–62 (1998).
7. Fedo, C. M., Myers, J. S. & Appel, P. W. U. Depositional setting and paleogeographic implications of earth's oldest supracrustal rocks, the >3.7 Ga Isua Greenstone belt, West Greenland. *Sedim. Geol.* **141**, 61–77 (2001).
8. Dymek, R. F. & Klein, C. Chemistry, petrology and origin of banded iron formation lithologies from the 3800 MA Isua supracrustal belt, West Greenland. *Precamb. Res.* **39**, 247–302 (1988).
9. Myers, J. S. Protoliths of the 3.8–3.7 Ga Isua greenstone belt, West Greenland. *Precamb. Res.* **105**, 129–141 (2001).
10. Nutman, A. P., Allaart, J. H., Bridgwater, D., Dimroth, E. & Rosing, M. Stratigraphic and geochemical evidence for the depositional environment of the early Archaean Isua supracrustal belt, southern West Greenland. *Precamb. Res.* **25**, 365–396 (1984).

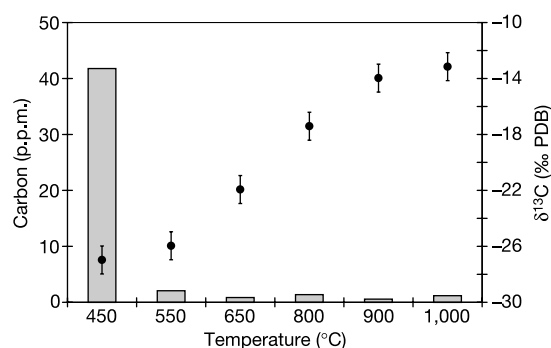


Figure 4 Stepped-heating combustion data for sample AL25. 1-h steps at 450, 550, 650, 800, 900 and 1,000 °C. Abundance of extracted reduced carbon (p.p.m.) and the isotopic composition ($\delta^{13}\text{C}$) are shown for each temperature step. The reduced carbon component that dominates in this sample is isotopically light ($\delta^{13}\text{C} = -28\text{‰}$) and comes off at low temperature.

11. Dimroth, E. in *Sedimentary Geology of the Highly Metamorphosed Precambrian Complexes* (ed. Sidorenko, A. V.) 16–27 (Nauka, Moscow, 1982).
12. Rose, N. M., Rosing, M. T. & Bridgwater, D. The origin of metacarbonate rocks in the Archaean Isua Supracrustal Belt, West Greenland. *Am. J. Sci.* **296**, 1004–1044 (1996).
13. Rosing, M. T., Rose, N., Bridgwater, D. & Thomsen, H. S. Earliest part of Earth's stratigraphic record: A reappraisal of the >3.7 Ga Isua (Greenland) supracrustal sequence. *Geology* **24**, 43–46 (1996).
14. Rosing, M. T. & Rose, N. M. The role of ultramafic rocks in regulating the concentrations of volatile and non-volatile components during deep crustal metamorphism. *Chem. Geol.* **108**, 187–200 (1993).
15. Hayes, J. M., Kaplan, I. R. & Wedekind, W. in *Earth's Earliest Biosphere, its Origin and Evolution* (ed. Schopf, J. W.) 93–134 (Princeton Univ. Press, New Jersey, 1983).
16. Oehler, D. Z. & Smith, J. W. Isotopic composition of reduced and oxidized carbon in Early Archaean rocks from Isua, Greenland. *Precamb. Res.* **5**, 221–228 (1977).
17. Perry, E. C. & Ahmad, S. N. Carbon isotope composition of graphite and carbonate minerals from 3.8 AE metamorphosed sediments, Isua, Greenland. *Earth Planet. Sci. Lett.* **36**, 280–284 (1977).
18. Ueno, Y., Yurimoto, H., Yoshioka, H., Komiya, T. & Maruyama, S. Ion microprobe analysis of graphite from ca. 3.8 Ga metasediments, Isua supracrustal belt, West Greenland: Relationship between metamorphism and carbon isotopic composition. *Geochim. Cosmochim. Acta* **66**, 1257–1268 (2002).
19. Abed, A. M. & Fakhouri, K. in *Phosphorite Research and Development* (eds Notholt, A. J. G. & Jarvis, I.) 193–203 (The Geological Society, London, 1990).
20. Appel, P. W. U. On the early Archaean Isua iron-formation, West Greenland. *Precamb. Res.* **11**, 73–87 (1980).
21. French, B. M. Stability relations of siderite (FeCO₃) in the system Fe-C-O. *Am. J. Sci.* **271**, 37–78 (1971).
22. van Zuilen, M. A. *et al.* Graphite and associating carbonates in early Archaean Isua supracrustal rocks, southern West Greenland. *Precamb. Res.* (submitted).
23. Chacko, T., Mayeda, T. K., Clayton, R. N. & Goldsmith, J. R. Oxygen and carbon isotope fractionations between CO₂ and calcite. *Geochim. Cosmochim. Acta* **55**, 2867–2882 (1991).
24. Boak, J. L. & Dymek, R. F. Metamorphism of the ca. 3800 Ma supracrustal rocks at Isua, West Greenland: implications for early Archaean crustal evolution. *Earth Planet. Sci. Lett.* **59**, 155–176 (1982).
25. Nagy, B. Porosity and permeability of the Early Precambrian Onverwacht chert: Origin of the hydrocarbon content. *Geochim. Cosmochim. Acta* **34**, 525–527 (1970).
26. Lepland, A., Arrhenius, G. & Cornell, D. Apatite in early Archaean Isua supracrustal rocks, southern West Greenland: its origin, association with graphite and potential as a biomarker. *Precambrian Res.* (in the press).
27. Fedo, C. M. & Whitehouse, M. J. Metasomatic origin of quartz-pyroxene rock, Akillia, Greenland, and implications for Earth's earliest life. *Science* **296**, 1448–1452 (2002).
28. Rosing, M. T. ¹³C-Depleted carbon microparticles in >3700-Ma sea-floor sedimentary rocks from West Greenland. *Science* **283**, 674–676 (1999).
29. Macpherson, C. G., Hilton, D. R., Newman, S. & Mathey, D. P. CO₂, ¹³C/¹²C and H₂O variability in natural basaltic glasses: A study comparing stepped heating and FTIR spectroscopic techniques. *Geochim. Cosmochim. Acta* **63**, 1805–1813 (1999).

Acknowledgements

We thank M. Wahlen and B. L. Deck for providing facilities for extraction of reduced carbon and subsequent isotopic measurement, D. R. Hilton for providing facilities for stepped-combustion extraction of reduced carbon, J. L. Teranes for measuring δ¹³C of carbonate phases, J. Finarelli for determination of cation composition of carbonate phases, and P. W. U. Appel for providing coordination and facilities for field work as part of the Isua Multidisciplinary Research Project. Support by the Marianne and Marcus Wallenberg Foundation (for A.L.) and NASA Exobiology is gratefully acknowledged. We thank L. P. Knauth, S. Moorbath and J. M. Hayes for their comments on this manuscript.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.A.v.Z. (e-mail: mvanzuil@ucsd.edu).

Cultivation of the ubiquitous SAR11 marine bacterioplankton clade

Michael S. Rappé, Stephanie A. Connon, Kevin L. Vergin & Stephen J. Giovannoni

Department of Microbiology, Oregon State University, Corvallis, Oregon 97331, USA

The α-proteobacterial lineage that contains SAR11 and related ribosomal RNA gene clones was among the first groups of organisms to be identified when cultivation-independent approaches based on rRNA gene cloning and sequencing were applied to survey microbial diversity in natural ecosystems¹. This group accounts for 26% of all ribosomal RNA genes that have

been identified in sea water and has been found in nearly every pelagic marine bacterioplankton community studied by these methods². The SAR11 clade represents a pervasive problem in microbiology: despite its ubiquity, it has defied cultivation efforts. Genetic evidence suggests that diverse uncultivated microbial taxa dominate most natural ecosystems^{3–5}, which has prompted widespread efforts to elucidate the geochemical activities of these organisms without the benefit of cultures for study^{6,7}. Here we report the isolation of representatives of the SAR11 clade. Eighteen cultures were initially obtained by means of high-throughput procedures for isolating cell cultures through the dilution of natural microbial communities into very low nutrient media. Eleven of these cultures have been successfully passaged and cryopreserved for future study. The volume of these cells, about 0.01 μm³, places them among the smallest free-living cells in culture.

In an effort to isolate some of the ubiquitous uncultivated Bacteria and Archaea that dominate marine bacterioplankton communities², we inoculated fresh Oregon coast seawater samples into microtitre dish wells by dilution, such that on average each well received 22 microbial cells. Media consisted of sterile Oregon coast sea water supplemented with either phosphate (as KH₂PO₄) and ammonium (as NH₄Cl), or phosphate, ammonium and a defined mixture of organic carbon compounds. The technique of isolating cells by dilution into sterilized natural waters or other dilute media has been used previously^{8,9}; it takes advantage of the fact that

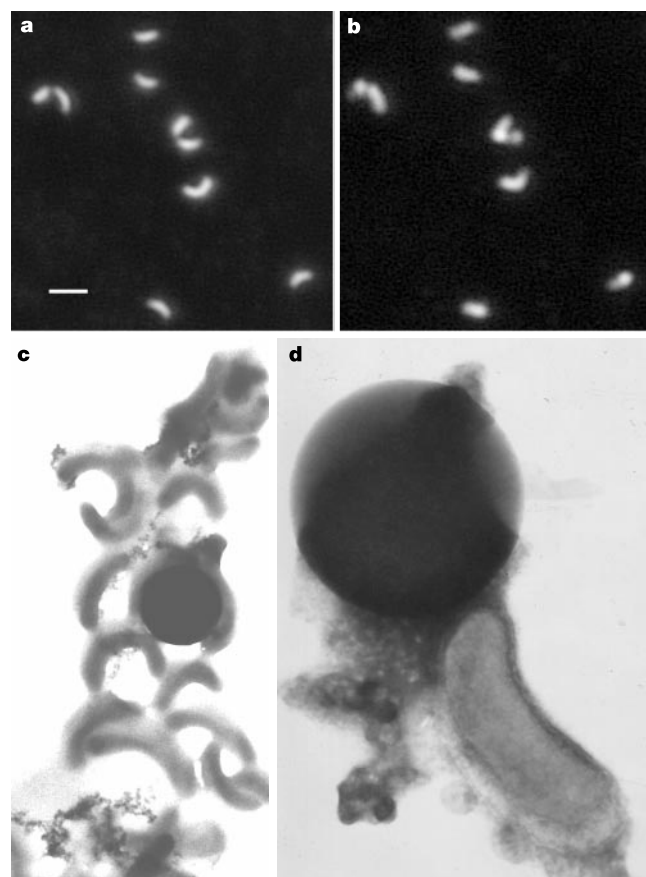


Figure 1 Photomicrographs of a culture of SAR11 clade isolate HTCC1062. **a, b**, Fluorescence images of cells in an identical field of view, stained with the DNA-specific dye DAPI (**a**) and after hybridization with four Cy3-labelled oligonucleotide probes targeting SAR11 cells (**b**). Scale bar (**a, b**), 1 μm. **c, d**, Transmission electron micrographs of strain HTCC1062. **c**, Shadowed cells with the typical SAR11 clade morphology. **d**, Negatively stained cell. The latex beads in **c** and **d** have a diameter of 0.514 μm.

substrate concentrations and cell numbers in natural waters are typically about three orders of magnitude less than in common laboratory media. The approach that we used to isolate members of the SAR11 clade was similar but involved modifications to a microtitre dish format and a newly developed procedure for making arrays of cell cultures on microscope slides coupled with fluorescence *in situ* hybridization (FISH)^{10,11}. These procedures were designed to increase the rate at which cultures could be obtained and identified. In this experiment, 288 extinction cultures of 1 ml were inoculated and screened.

After incubation at 15 °C for 23 d either in the dark or under a 14 h/10 h light/dark cycle (irradiance $\approx 400 \mu\text{mol m}^{-2} \text{s}^{-1}$), culture wells were tested for replicating cells by arraying small culture volumes on polycarbonate membranes. The limit of detection with this method is roughly 2×10^3 cells per ml, and it requires only 100 μl of culture (~ 200 cells). Arrays were stained with the DNA-binding dye 4',6-diamidino-2-phenylindole (DAPI)¹² and positive wells were screened further by FISH with a set of four fluorescently labelled oligonucleotide probes specific for SAR11 clade 16S rRNA (Fig. 1). Ten cultures consisting of 100% SAR11 clade cells and eight mixed cultures containing about 10–90% of SAR11 clade cells were identified preliminarily. Cultures were obtained in microtitre plates incubated in the dark or under a light/dark cycle, and in both the medium containing only natural organic carbon and the medium supplemented with a defined mixture of carbon compounds. Eleven isolates were propagated and cryopreserved for future study.

Phylogenetic relationships of the isolates were investigated by a combination of 16S rRNA gene and 16S–23S rDNA intergenic spacer sequence analysis. Ribosomal RNA nucleotide sequences of 414–608 bases, obtained from the 3' end of the 16S rRNA gene of all 11 isolates, were identical. Intergenic spacer sequences (415–417 bases) indicated, however, that there were three genetically distinct

lineages among the strains. Strains HTCC1002, 1016, 1025, 1056, 1057 and 1061 differed from HTCC1004, 1013, 1040 and 1062 at ten sequence positions and by an indel of two nucleotides. An eleventh isolate (HTCC1051) differed from the latter group at two nucleotide positions. The two principal sequence groups contained strains that were originally isolated from dark and light/dark incubations, as well as from both media types. Complete 16S rRNA gene sequences obtained from representatives of the two main intergenic spacer groups (HTCC1002 and HTCC1062) were found to differ by a single base and were more than 99% similar to over 30 published nucleotide sequences from rRNA-gene-containing clones recovered from sea water. These include nearly full-length rDNA clones obtained from the Arctic Ocean^{13,14}, from the coast of Plymouth, UK, in the Atlantic Ocean¹⁵ (Fig. 2) and from the Santa Barbara Channel¹⁶ (data not shown).

Comparative sequence analysis of 16S rRNA genes showed that SAR11 and its relatives are a deeply branching cluster of the α -subclass of Proteobacteria. Members of this group show less than 82% sequence similarity to cultivated members of the α -Proteobacteria. Since their original discovery in the Sargasso Sea, members of the SAR11 clade have been recovered in every 16S rRNA gene clone library that has been constructed with universal or bacterial polymerase chain reaction (PCR) primers from marine prokaryotic plankton samples, including coastal and near-shore waters^{16,17} and seawater samples from depths up to 3,000 m (ref. 18). Relatives of the SAR11 clade have even been detected in freshwater lakes¹⁹.

Growth rates for the 11 SAR11 clade isolates replicating at 15 °C in sterile Oregon coast seawater supplemented with 0.1 μM phosphate, 1.0 μM ammonium and a defined mixture of organic carbon compounds ranged from 0.40 to 0.58 d^{-1} . Although this rate of cell division is low in comparison to values typical of cultivated bacteria, it is not dissimilar to the measured growth rates of marine bacterioplankton communities in nature, which vary from 0.05 to 0.3 d^{-1} (ref. 20). All of the isolates produced a logistic growth curve (Fig. 3). In subsequent experiments with strain HTCC1062, which was originally isolated in seawater media supplemented with the defined organic carbon compound and vitamin mixtures, removal of these amendments did not negatively affect growth rate or maximum cell abundance (Fig. 3). However, the addition of dilute proteose peptone (0.001%) inhibited growth (Fig. 3).

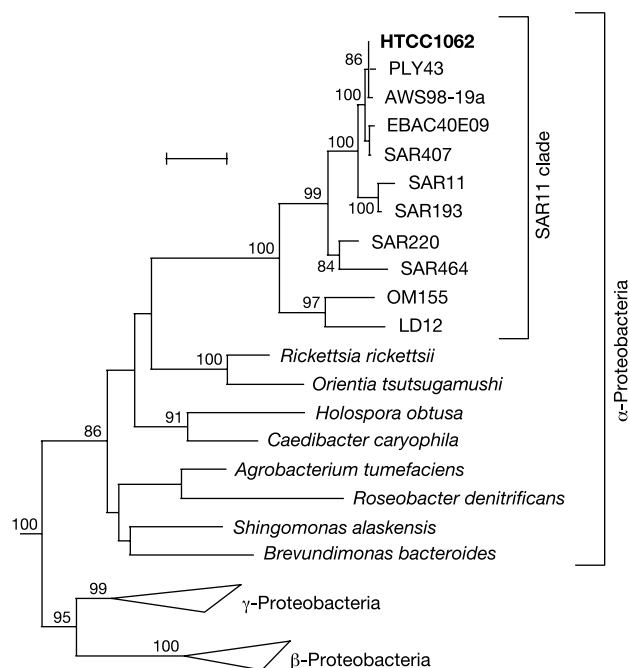


Figure 2 Phylogenetic relationships between strain HTCC1062 and representatives of the SAR11 clade and α -Proteobacteria inferred from 16S rRNA gene sequence comparisons. The Gram-positive bacteria *Bacillus subtilis* and *Marinococcus halophilus* were used as outgroups. Bootstrap proportions over 70% that supported the branching order are shown. Scale bar corresponds to 0.05 substitutions per nucleotide position. Also included in the analysis were the γ -Proteobacteria *Alteromonas macleodii* and *Marinobacter hydrocarbonoclasticus*, and the β -Proteobacteria *Methylophilus methylotrophus* and *Polynucleobacter necessarius*.

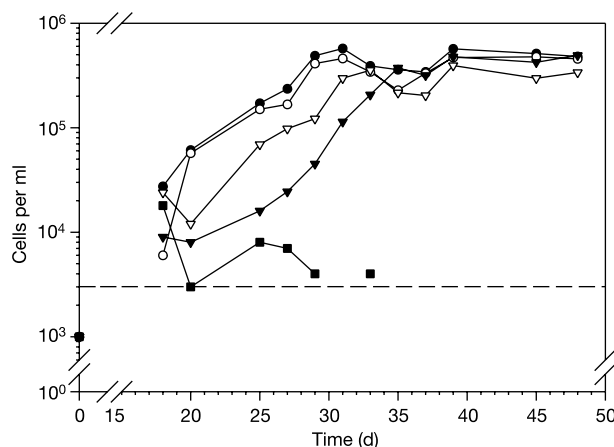


Figure 3 Growth of strain HTCC1062 in Oregon coast seawater media. Media consisted of sterile sea water supplemented with 1.0 μM NH_4Cl and 0.1 μM KH_2PO_4 (filled circles), 1.0 μM NH_4Cl , 0.1 μM KH_2PO_4 , and mixed carbon (open circles), 1.0 μM NH_4Cl , 0.1 μM KH_2PO_4 , and Va vitamins (filled triangles), 1.0 μM NH_4Cl , 0.1 μM KH_2PO_4 , mixed carbon and Va vitamins (open triangles), 1.0 μM NH_4Cl , 0.1 μM KH_2PO_4 , mixed carbon, Va vitamins and 0.001% (w/v) proteose peptone (filled squares). For all cultures, cell counts attempted on days 7 and 12 were below the limit of detection (dotted line, 3,000 cells per ml), as were counts on day 31 and after day 33 for the culture containing proteose peptone. The point at day 0 is the inoculum density.

The maximum cell densities reached by the SAR11 clade isolates differed between seawater samples collected on different dates and locations in the Pacific Ocean off the Oregon coast, regardless of the addition of organic carbon. Overall, the maximum carrying capacity of the different seawater batches varied from 2.5×10^5 cells per ml to 3.5×10^6 cells per ml. These results suggest that natural factors present in seawater control the cultured SAR11 populations. Without nutrient amendments, some sterile sea water batches supported the growth of these isolates up to cell densities of 2.5×10^5 cells per ml. Other Oregon coast seawater would only promote the growth of these isolates when at least $1.0 \mu\text{M}$ ammonium and $0.1 \mu\text{M}$ phosphate were added. The growth of these cells did not seem to be affected by light. Overall, these observations have important implications for oceanographic research because they suggest that cultured SAR11 clade cells growing in sea water can be used to identify chemical factors that control their natural abundance in the environment.

SAR11 clade cells are crescent-shaped (vibrioid), $0.37\text{--}0.89 \mu\text{m}$ in length, and have an average cell diameter of $0.12\text{--}0.20 \mu\text{m}$ (Fig. 1). The size and morphology of the isolates was identical to cells of the SAR11 clade that have been observed in natural populations by fluorescence *in situ* hybridization. Size estimates made by transmission electron microscopy on a culture of HTCC1062 cells indicate that it is one of the smallest free-living and replicating pure cultures of cells known (Fig. 1). The size differences of these estimates arise from the natural variation in cell length that is associated with cell division, and uncertainty about whether a lightly staining structure observed on the outwardly curved lateral surface of cells is enclosed by membrane (Fig. 1). In addition, the cells were measured after fixation in glutaraldehyde, which may have caused shrinkage. For a cell of $0.4 \times 0.2 \mu\text{m}^2$, which is a conservative estimate of the size of an HTCC1062 cell emerging from division, the cell volume is about $0.01 \mu\text{m}^3$. The small size of the SAR11 clade isolates is noteworthy, particularly considering that this organism is well adapted to a life of autonomous replication. The genome size of strain HTCC1062 was estimated to be 1.54 Mb by pulsed-field gel electrophoresis.

The success of this cultivation approach with members of the SAR11 clade may be attributed to several factors, including the use of pristine sea water as a medium, the relatively large number of cultures screened, and the low detection threshold of the procedure. Access to members of the SAR11 clade in culture will provide an unusual opportunity for genome sequence analysis of an organism that has global biogeochemical significance and can be manipulated in culture. It may also provide insight into the adaptations of cells to very low nutrient systems. Because of their extraordinarily small size, the study of these isolates should also refine our understanding of the minimal macromolecular machinery required for autonomous cellular replication. On the basis of the preliminary characteristics presented here, we propose the status '*Candidatus Pelagibacter ubique*' gen. nov., sp. nov., for this taxon. □

Methods

Sample collection and culturing

A water sample was collected in a 2-l Niskin bottle fitted with a Teflon-coated spring from a depth of 10 m at a station 27.6 km off the coast of Oregon on the Newport Hydrographic line ($44^\circ 39.1' \text{ N}$, $124^\circ 24.7' \text{ W}$). The bottle was immediately stored in a cooler with cold packs to maintain a temperature similar to surface water at this station (12.0°C) until further laboratory processing (3 h from sample collection). The primary cultivation experiments used two types of media: filtered ($0.2 \mu\text{m}$), autoclaved, CO_2 - and air-sparged coastal Pacific Ocean sea water amended with $1.0 \mu\text{M}$ NH_4Cl and $0.1 \mu\text{M}$ KH_2PO_4 ; and an ammonium- and phosphate-enriched medium amended with 0.001% (w/v) D-glucose, D-ribose, succinic acid, pyruvic acid, glycerol, N-acetyl glucosamine, 0.002% (v/v) ethanol and Va vitamin mix at a 10^{-4} dilution of stock²¹. Growth curve experiments used the same nutrient concentrations.

Fluorescent microscopy

For the identification of SAR11 clade-containing cultures by FISH, we fixed 100–200 μl from each culture well with $0.2\text{-}\mu\text{m}$ -filtered paraformaldehyde (2% final concentration)

and filtered it onto $0.2\text{-}\mu\text{m}$ polycarbonate membrane filters. FISH was done essentially as described²², with the following modifications (R. Morris, unpublished data). Hybridization reactions were carried out at a temperature of 35°C for 3–12 h in hybridization buffer consisting of 900 mM NaCl, 20 mM Tris (pH 7.4), 0.01% (w/v) SDS, 15% formamide and four Cy3-labelled oligonucleotide probes targeting SAR11 clade ribosomal RNA at a final concentration of $2 \text{ ng } \mu\text{l}^{-1}$ each (SAR11-152R, 5'-ATTAGCAC AAGTTTCCYCGTGT-3'; SAR11-441R, 5'-TACAGTCATTTCTTCCCCGAC-3'; SAR11-542R, 5'-TCCGAACCTACGCTAGGTC-3'; SAR11-732R, 5'-GTCAGTAATGATCC AGAAAGYTG-3'). We achieved optimal hybridization stringency by washing the membranes for two 15-min intervals at 55°C in 150 mM NaCl, 20 mM Tris (pH 7.4), 6 mM EDTA and 0.01% SDS. After mounting filters in Citifluor (Ted Pella), Cy3-positive and DAPI-positive cells were counted for each field of view²³ using a Leica DMRB epifluorescence microscope equipped with a Hamamatsu ORCA-ER CCD digital camera, filter sets appropriate for Cy3 and DAPI, and Scanalytics IPlab v3.5.5 scientific imaging software.

Electron microscopy

Cells were initially concentrated by centrifugation or Vivascience Vivaspin 500 ultrafiltration concentrators and resuspended in sterile sea water containing 0.5% glutaraldehyde. After centrifugation of concentrated cell suspensions onto Formvar-coated copper grids, cells were prepared for imaging by either negative staining with 2% uranyl acetate (pH 4.0), or shadowing with gold/platinum. We included fluoresbrite beads (diameter, $0.514 \pm 0.015 \mu\text{m}$) in the preparations as an internal size standard.

Genome size

Concentrated cells of strain HTCC1062 were embedded in agarose, and individual cell plugs were digested with the homing endonuclease I-CeuI or the restriction endonucleases *NotI* or *SfiI*. We carried out electrophoresis with a CHEF DRII unit in 1% agarose for 18 h at 6 V cm^{-1} , with an initial switch time of 50 s and a final switch time of 90 s. Genome size was measured by comparison to a yeast chromosome PFG marker. The I-CeuI endonuclease apparently linearized the chromosome, resulting in a single band. The restriction endonucleases *NotI* and *SfiI* did not cut the genomic DNA of strain HTCC1062.

Phylogenetic analysis

For small-volume ($\sim 1 \text{ ml}$) cultures, genomic DNA was isolated from 200 μl of culture using a DNeasy Tissue kit (Qiagen) after two freeze–thaw cycles. For large volume cultures ($\sim 100 \text{ ml}$), cells were filtered onto $0.1\text{-}\mu\text{m}$ pore size Supor-100 polysulfone filters and the DNA was extracted as described²⁴. Crude nucleic acid extracts were purified further with a DNeasy Tissue kit. Owing to the small amount of starting material, a semi-nested PCR reaction was sometimes required to obtain sufficient product for further characterization. We characterized the 3' end of the 16S rRNA gene and the intergenic transcribed spacer in a single reaction by PCR amplification with the 16S rRNA gene primer 926F and the 23S rRNA gene primer 23S-117R (ref. 25), or in semi-nested reactions using the forward primer corresponding to the SAR11-542R oligonucleotide followed by PCR amplification with the 926F primer (both paired with 23S-117R). Some semi-nested reactions also used 926F followed by 1100F as forward primers (both paired with 23S-117R). For nearly complete 16S rRNA gene sequencing, genes were initially amplified in a PCR reaction using the primers 27F and 1492R, followed by a second reaction using 27F and 1406R. PCR products were cleaned with a QiaQuick PCR purification column (Qiagen) and sequenced on an ABI 377 or ABI 3100 automated sequencer.

We aligned sequences against those in a database of over 11,500 SSU rDNA sequences maintained with the ARB software package²⁶. We carried out phylogenetic analyses with the program PAUP* 4.0 beta 8 (ref. 27), and included 951 unambiguously aligned nucleotide positions. The tree topology was inferred by maximum likelihood with a heuristic search and tree bisection-reconnection (TBR) branch swapping, a transition/transversion ratio that was estimated from the data (1.426) and nucleotide frequencies that were estimated from the data. We determined bootstrap proportions from 1,000 resamplings using evolutionary distances calculated with the Kimura 2-parameter model for nucleotide change, a transition/transversion ratio that was estimated from the data, and neighbour joining.

Received 11 April; accepted 22 May 2002; doi:10.1038/nature00917.

- Giovannoni, S. J., Britschgi, T. B., Moyer, C. L. & Field, K. G. Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**, 60–63 (1990).
- Giovannoni, S. & Rappé, M. in *Microbial Ecology of the Oceans* (ed. Kirchman, D. L.) 47–84 (Wiley, New York, 2000).
- Hugenholtz, P., Goebel, B. M. & Pace, N. R. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J. Bacteriol.* **180**, 4765–4774 (1998).
- Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740 (1997).
- Ward, D. M., Bateson, M. M., Weller, R. & Ruff-Roberts, A. L. Ribosomal RNA analysis of microorganisms as they occur in nature. *Adv. Microb. Ecol.* **12**, 219–286 (1992).
- Beja, O., Spudich, E. N., Spudich, J. L., Leclerc, M. & DeLong, E. F. Proterorhodopsin phototrophy in the ocean. *Nature* **411**, 786–789 (2001).
- Stein, J. L., Marsh, T. L., Wu, K. Y., Shizuya, H. & DeLong, E. F. Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J. Bacteriol.* **178**, 591–599 (1996).
- Schut, F. et al. Isolation of typical marine bacteria by dilution culture: growth, maintenance, and characteristics of isolates under laboratory conditions. *Appl. Environ. Microbiol.* **59**, 2150–2160 (1993).
- Button, D. K., Schut, F., Quang, P., Martin, R. & Robertson, B. Viability and isolation of marine bacteria by dilution culture: theory, procedures, and initial results. *Appl. Environ. Microbiol.* **59**, 881–891 (1993).
- Giovannoni, S. J., DeLong, E. F., Olsen, G. J. & Pace, N. R. Phylogenetic group-specific oligodeoxynucleotide probes for identification of single microbial cells. *J. Bacteriol.* **170**, 720–726 (1988).

11. Amann, R. I., Krumholz, L. & Stahl, D. A. Fluorescent-oligonucleotide probing of whole cells for determinative, phylogenetic, and environmental studies in microbiology. *J. Bacteriol.* **172**, 762–770 (1990).
12. Porter, K. G. & Feig, Y. S. The use of DAPI for identifying and counting aquatic microflora. *Limnol. Oceanogr.* **25**, 943–948 (1980).
13. Bano, N. & Hollibaugh, J. T. Phylogenetic composition of bacterioplankton assemblages from the Arctic Ocean. *Appl. Environ. Microbiol.* **68**, 505–518 (2002).
14. Yager, P. L. *et al.* Dynamic bacterial and viral response to an algal bloom at subzero temperatures. *Limnol. Oceanogr.* **46**, 790–801 (2001).
15. Rochelle, P. A., *et al.* in *Nucleic Acids in the Environment* (eds Trevors, J. T. & van Elsland, J. D.) 219–239 (Springer, Berlin, 1995).
16. DeLong, E. F., Franks, D. G. & Alldredge, A. L. Phylogenetic diversity of aggregate-attached vs free-living marine bacterial assemblages. *Limnol. Oceanogr.* **38**, 924–934 (1993).
17. Rappé, M. S., Kemp, P. F. & Giovannoni, S. J. Phylogenetic diversity of marine coastal picoplankton 16S rRNA genes cloned from the continental shelf off Cape Hatteras, North Carolina. *Limnol. Oceanogr.* **42**, 811–826 (1997).
18. Fuhrman, J. A. & Davis, A. A. Widespread *Archaea* and novel *Bacteria* from the deep sea as shown by 16S rRNA gene sequences. *Mar. Ecol. Prog. Ser.* **150**, 275–285 (1997).
19. Bahr, M., Hobbie, J. E. & Sogin, M. L. Bacterial diversity in an arctic lake: a freshwater SAR11 cluster. *Aquat. Microb. Ecol.* **11**, 271–277 (1996).
20. Ducklow, H. in *Microbial Ecology of the Oceans* (ed. Kirchman, D. L.) 85–120 (Wiley, New York, 2000).
21. Davis, H. C. & Guillard, R. R. L. Relative value of ten genera of micro-organisms as food for oyster and clam larvae. *USFWS Fish Bull.* **58**, 293–304 (1958).
22. Glöckner, F. O. *et al.* An *in situ* hybridization protocol for detection and identification of planktonic bacteria. *System. Appl. Microbiol.* **19**, 403–406 (1996).
23. Hicks, R. E., Amann, R. I. & Stahl, D. A. Dual staining of natural bacterioplankton with 4',6-diamidino-2-phenylindole and fluorescent oligonucleotide probes targeting kingdom-level 16S rRNA sequences. *Appl. Environ. Microbiol.* **58**, 2158–2163 (1992).
24. Giovannoni, S. J., DeLong, E. F., Schmidt, T. M. & Pace, N. R. Tangential flow filtration and preliminary phylogenetic analysis of marine picoplankton. *Appl. Environ. Microbiol.* **56**, 2572–2575 (1990).
25. García-Martínez, J. & Rodríguez-Valera, F. Microdiversity of uncultured marine prokaryotes: the SAR11 cluster and the marine *Archaea* of Group I. *Mol. Ecol.* **9**, 935–948 (2000).
26. Ludwig, W. *et al.* Bacterial phylogeny based on comparative sequence analysis. *Electrophoresis* **19**, 554–568 (1998).
27. Swofford, D. *PAUP* 4.0* (Sinauer Associates, Sunderland, Massachusetts, 2000).

Acknowledgements

We thank R. Morris and C. Alexander for technical assistance; A. Soeldner and M. Nesson for electron microscopy expertise; W. Peterson, L. Feinberg and the US GLOBEC Program for CTD data; and the crew of the RV *Elakha*. This research was supported by Diversa Corporation and the National Science Foundation.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.J.G. (e-mail: steve.giovannoni@orst.edu). The sequences have been deposited in GenBank under accession numbers AF510191 to AF510193.

An unexpected specialization for horizontal disparity in primate primary visual cortex

B. G. Cumming

Laboratory of Sensorimotor Research, National Eye Institute, National Institutes of Health, Bethesda, Maryland 20982, USA

The horizontal separation of the eyes means that objects nearer or farther than the fixation point project to different locations on the two retinæ, differing principally in their horizontal coordinates (horizontal binocular disparity). Disparity-selective neurons have generally been studied with disparities applied in only one direction¹ (often horizontal), which cannot determine whether the encoding is specialized for processing disparities along the horizontal axis. It is therefore unclear if disparity selectivity represents a specialization for naturally occurring disparities. I used random dot stereograms to study disparity-

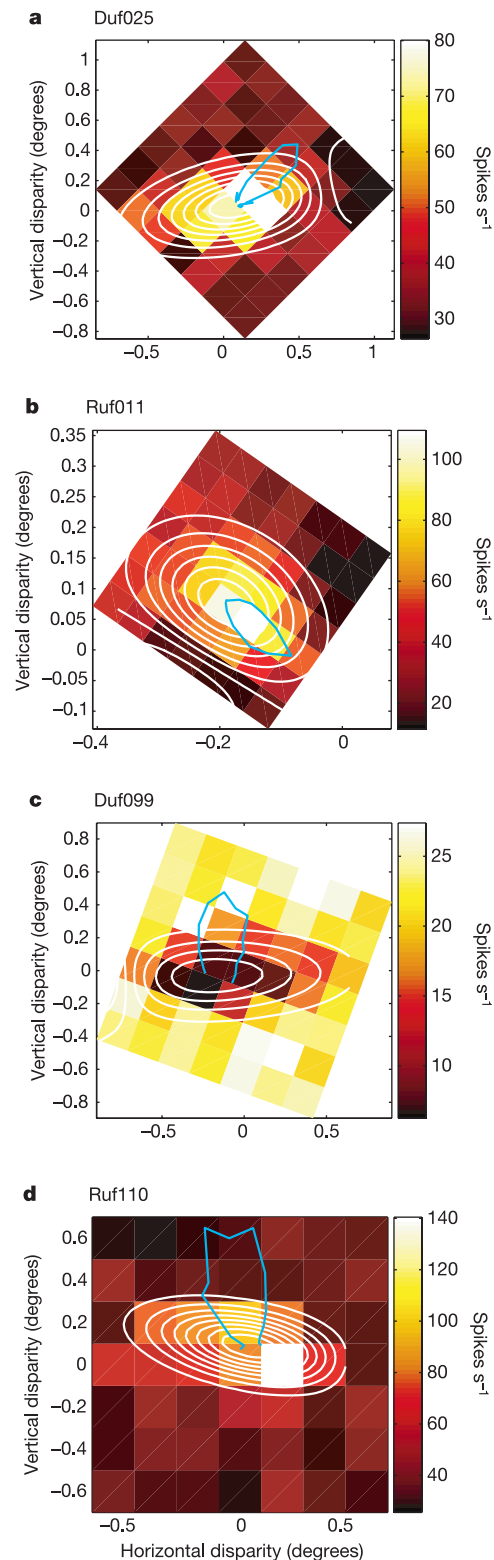


Figure 1 Responses of four neurons to combinations of vertical and horizontal disparities. The white contour lines show the fitted two-dimensional Gabor function (see Methods). A polar plot of the monocular orientation tuning (stimuli spanning 180°) for each cell is superimposed (blue line, origin at the peak of the fitted Gabor). **a**, Despite an oblique orientation preference, cell 025 from monkey Duf (Duf025) produces a horizontally oriented disparity response. **b**, Ruf011 shows an obliquely oriented disparity response, matching the receptive-field orientation. **c**, **d**, Despite preferred orientations close to vertical, the disparity response surfaces for both Ruf110 and Duf099 are horizontally elongated. Duf099 illustrates a 'tuned inhibitory' response type. The s.e.m. for each data point was small relative to the mean (mean ratio 0.11).

selective neurons from the primary visual cortex (V1) of awake fixating monkeys. Many combinations of vertical and horizontal disparity were used, characterizing the surface of responses as a function of two-dimensional disparity. Here I report that the response surface usually showed elongation along the horizontal disparity axis, despite the isotropic stimulus. Thus these neurons modulated their firing rate over a wider range of horizontal disparity than vertical disparity. This demonstrates that disparity-selective cells are specialized for processing horizontal disparity, and that existing models^{2,3} of disparity selectivity require substantial revision.

Points outside the fixation plane project to different locations on the two retinae. As the eyes are separated horizontally, these locations are related by a displacement along an axis very close to horizontal (horizontal disparity). There may also be a small vertical component⁴, but this is determined by eye position rather than the three-dimensional properties of the scene. Consequently, in central vision, the variation in disparity occurs along an axis that is very close to horizontal. Although there is a need for some information about vertical disparity to maintain alignment of the eyes and estimate viewing distance⁵, this does not require many finely spaced measures across the retina. Conversely, finely spaced measures of horizontal disparity are required to detect variations in depth within the scene.

It therefore seems natural to assume that neurons selective for binocular disparity encode horizontal disparities, but in principle they may signal disparities along other axes. Indeed, current understanding of the mechanism by which V1 receptive fields are constructed suggests that, depending on receptive-field orientation, different neurons signal disparities along different axes^{1,6}. However, disparity-selective neurons have generally been tested with disparities applied along only one axis, which cannot reveal whether or not the neuronal responses are specialized to exploit the horizontal bias of naturally occurring disparities. In order to do this, it is necessary to explore responses of single neurons to disparities applied along several axes, using a visual stimulus that is isotropic (such as random dot stereograms, RDS).

Sixty disparity-selective neurons from V1 of two awake fixating monkeys were stimulated with combinations of vertical and horizontal disparity in RDS. Example responses are shown in Fig. 1. In all cases, large horizontal or vertical disparities yield a baseline firing rate, reflecting the response to uncorrelated dots within the receptive field. A set of disparity combinations elicited responses different from this baseline, occupying a region that was approximately elliptical. For most neurons, the long axis of this ellipse was near horizontal—they modulated their firing rate over a wider range of

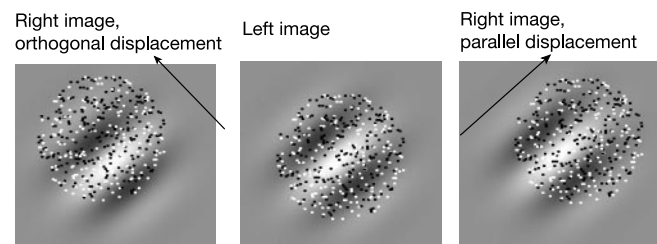


Figure 2 Relationship between disparity tuning and receptive-field orientation, for simple models¹. The receptive field (RF), depicted by the grey-level image, is selective for oblique contours. Superimposed is a random dot pattern like those used for the experiments (but this example has fewer dots, so that the RF is not obscured, and is smaller relative to the RF). Displacement of the dot patch orthogonal to the RF (left) should produce a stronger change in activity than a similar displacement parallel to the RF (right). The orientation of the disparity response surface should therefore match the monocular preferred orientation.

horizontal disparities than vertical disparities.

Figure 1b illustrates a less common pattern, where the response surface is not oriented horizontally. This might readily be explained by the properties of the monocular receptive field, as V1 neurons stimulated with extended contours show selectivity for the axis of orientation. If disparity selectivity is generated after this orientation-selective stage, simple models (like the binocular energy model^{2,3}) predict a correlation between monocular orientation preference and the orientation of the disparity response surface (Fig. 2). If the orientation selectivity is the result of an oriented linear filter applied to the image, any simple disparity measure (for example, cross-correlation between the filtered images) applied after the filter will reflect the filter orientation. To examine this relationship, the monocular orientation preference of each neuron was measured independently with sinusoidal luminance gratings. Figure 3 compares preferred (monocular) orientation with the orientation of the disparity tuning surface for each neuron. Many points lie close to the identity line, in accordance with simple models. However, the majority show a near horizontal orientation of their disparity tuned responses, regardless of the monocular preferred orientation. This bias is highly significant ($P \ll 0.001$, Rayleigh test), and is so strong that it obscures any correlation between preferred orientation and the orientation of the disparity responses across the population (T-monotone association 0.004, $P > 0.1$).

These horizontally oriented response surfaces represent a specialization for horizontal disparity in the sense that these neurons modulate their responses over a wider range of horizontal disparities than vertical disparities. The ability of the population to encode disparity also depends on differences between neurons, which could enable the population to encode a larger range than any individual

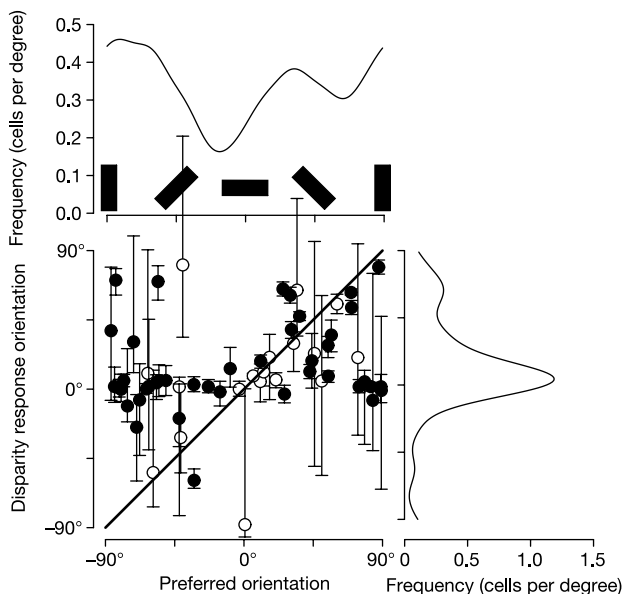


Figure 3 Orientation of the disparity tuned response as a function of monocular orientation preference (dominant eye). Bottom scatter-plot for 60 neurons (32 from monkey Duf, 28 from monkey Ruf). Error bars show 95% confidence intervals. The orientation of the disparity response surface tends to be near zero (horizontal), regardless of the monocular orientation preference. Data points for neurons in which the two orientations are significantly different ($P < 0.05$, by resampling) are filled. The diagonal line shows the identity function. Each datum here is based on a mean of 845 stimulus presentations. The smoothed frequency distribution on the right shows a strong peak very close to zero (at 4°). Top, the frequency distribution for monocular orientations (above) shows a slight (but insignificant: $P > 0.05$, Rayleigh test) bias, which is away from zero.

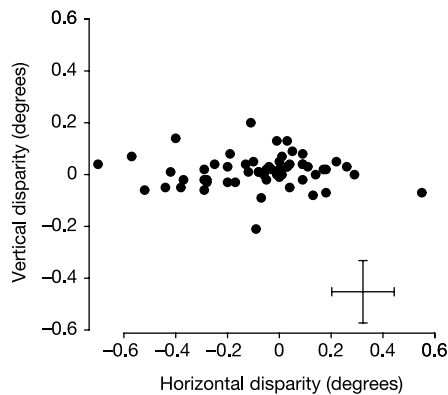


Figure 4 For each neuron, the preferred disparity identifies the disparity combination yielding the strongest response. The horizontal and vertical components of this preferred disparity are plotted here for each cell. The cross at the lower right shows the median value of the 95% confidence interval for each parameter. The horizontal components are more widely scattered than the vertical components, allowing the population to encode a wider range of horizontal disparities than vertical disparities.

neuron. The disparity that produced the strongest response was therefore examined for each neuron. This single point on the surface has horizontal and vertical disparity components. Differences between neurons in this preferred disparity were much greater for the horizontal than vertical components (Fig. 4, ratio of s.d.s 3.6, $P \ll 0.001$, F -test). A similar observation was made in one of the earliest studies of disparity tuning⁷, but a subsequent debate^{8,9} raised doubts about its validity. However, all of these early studies used bar stimuli, where the orientation of the stimulus strongly influences the two-dimensional disparity tuning¹. A few subsequent studies have used non-oriented stimuli^{10,11}, but did not present data on

receptive-field orientation.

Both features of the data (the horizontal orientation of the disparity surface, and the greater horizontal scatter) indicate a specialization for encoding a wide range of horizontal disparities. The same features of the data make individual neurons, and the population as a whole, more sensitive to small changes in vertical than horizontal disparity (this can be appreciated from the spacing of the contour lines in Fig. 1). The remarkable finding is that neurons with different preferred orientation all exhibit the same anisotropy. A different anisotropy has been demonstrated in anaesthetized cats—vertically oriented simple cells show greater interocular differences in receptive-field structure (phase) than those with horizontal orientations¹². I show here that the receptive-field orientation is not straightforwardly related to the direction of disparity encoding, so further work is required to understand the relationship between such monocular measures of receptive-field structure and disparity selectivity.

Considerable revision to the previously very successful energy model of disparity selectivity³ will be required to explain these data. One possibility is that end- or side-stopping¹³ precedes binocular combination, and hence partially determines the shape of the disparity-tuned response. (End-stopping might then prevent a neuron with a vertically oriented receptive field from responding to large vertical disparities.) An alternative is that these responses are constructed from many binocular subunits, differing from one another in their horizontal position disparity, but all having similar vertical position disparities. If the differences are large enough, the oriented structure of individual subunits can be obscured in the sum of all subunits. Although earlier observations also indicated a need to modify the model^{1,14–16}, none of those required such an extreme revision.

Both the properties of single neurons, and the differences between neurons in the population, serve to maximize the range of horizontal disparities encoded by disparity-selective neurons in V1. This demonstrates that the properties of these neurons reflect

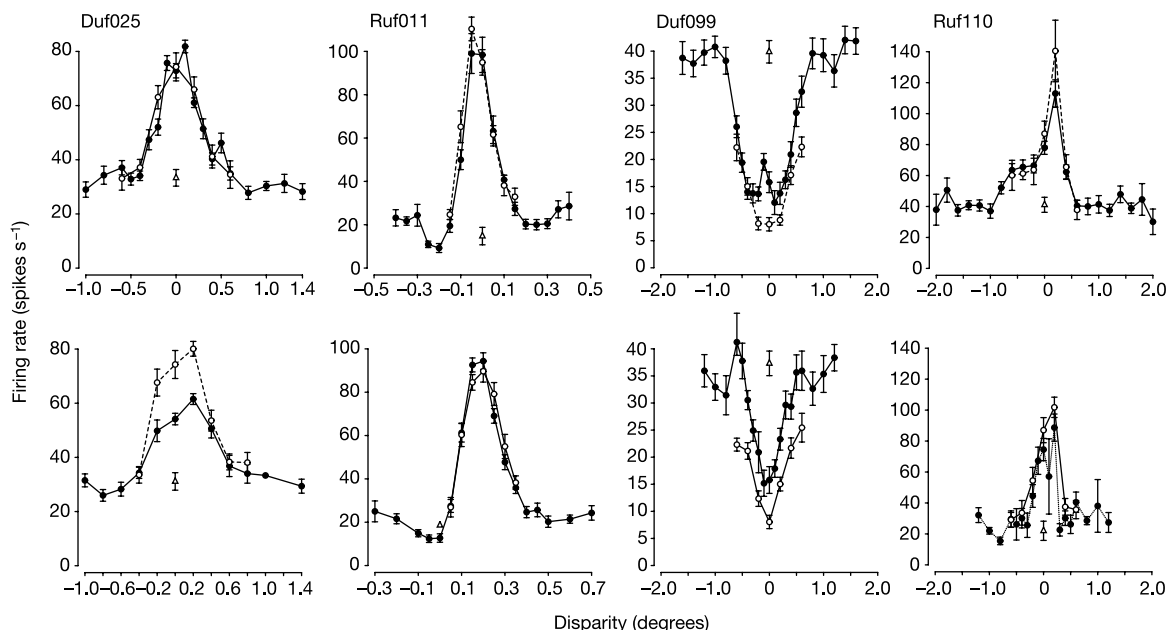


Figure 5 In order to select the disparity combinations used, a wide range of disparities were first explored along each of two axes (solid symbols). These axes were orthogonal to the receptive field (RF; top row) or parallel to the RF (bottom row). Each open triangle shows the response to binocularly uncorrelated dots. The open circles show the

comparable rows of data extracted from the surface plots in Fig. 1, for comparison. As these trials were not interleaved, there are sometimes differences in the absolute response rates. Nonetheless, it is clear that the response surfaces in Fig. 1 cover the important range of disparities for these neurons.

the anisotropy of naturally occurring disparities. In the light of recent evidence indicating that activity in V1 disparity-selective neurons does not directly support stereopsis^{14,17–19}, it is especially important to know if they show any specialization that is suited to a role in depth perception. The present study provides the strongest such evidence to date. □

Methods

Overview

A detailed description of the general methods has been published elsewhere^{17,20}. Briefly, extracellular single-unit recordings were made in the primary visual cortex of two alert monkeys (*Macaca mulatta*), trained to maintain fixation with both eyes. Stimuli were presented on two EIZO Flexscan F980 monitors (mean luminance 41.1 cd m⁻², contrast 99%), viewed via a mirror haploscope. The viewing distance was 89 cm, where each pixel subtended 0.019°. The positions of both eyes, along with the spike waveforms, were stored to disk for subsequent analysis. All protocols were approved by the Institute Animal Care and Use Committee, and complied with Public Health Service policy on the humane care and use of laboratory animals.

Experimental details

Stimuli were generated on a Silicon Graphics Octane workstation. Sinusoidal luminance gratings presented monocularly to each eye were used to determine the preferred orientation, spatial frequency, and minimum response field (MRF) of each neuron. Quantitative determination of the preferred orientation used a gaussian fit to data from at least nine orientations spanning 180°. MRFs were located at eccentricities between 2° and 9°. Dynamic RDS (generally 3° diameter) were centred over the MRF. Disparity-selective neurons were then tested with combinations of vertical and horizontal disparity, producing disparity displacements both parallel to and orthogonal to the preferred orientation. At least 49 combinations (seven parallel times seven orthogonal) were used. Stimuli were presented for 400 ms, and repeated a minimum of 5 times each (mean 13 repetitions). The range of disparities applied was determined by first measuring responses to a wide range of disparities (usually 15 disparities spanning $\pm 1.2^\circ$, but this range was extended if there was evident modulation at either extreme) applied along an axis orthogonal to the receptive-field orientation. A similar range was then tested along an axis parallel to the receptive field. (Significant tuning in these preliminary tests was a criterion for entry into the study). The two-dimensional grid was then chosen so as to cover the region with the strongest modulation. Figure 5 shows the relationship between these initial measures and those taken from the two-dimensional grid for the four cells illustrated in Fig. 1.

Data analysis

The mean firing rate as a function of the disparity combination, $f(x, y)$, was fitted with a single two-dimensional Gabor function:

$$f(x, y) = A \exp[(y - Y)^2 / 2\sigma_y^2] \exp[(x - X)^2 / 2\sigma_x^2] \cos[2\pi\omega(x - X) + \phi] + B$$

by nonlinear regression, where A , ω and ϕ are the amplitude, spatial frequency and phase of the cosine component, σ_x and σ_y are the standard deviations in orthogonal directions, X and Y are position offsets and B is the baseline firing rate. The two variables x and y were related to the horizontal and vertical disparity displacements by a single rotation through an angle θ , which was a free parameter in the fit. Simply using the fitted orientation of the Gabor function did not always correctly identify the orientation of the disparity surface. (When the surface is close to gaussian, two equivalent fits are possible in which the value of θ differs by 90°. Note that the Gabor still provides an excellent description, as a gaussian is a form of Gabor.) The fitted Gabor was therefore used only as a description of the data, from which a preferred disparity and orientation were determined. The preferred disparity was defined as the point that produced the largest deviation from the response to binocularly uncorrelated dots (maximum interaction position²⁰). This identifies the trough for TI cells, which are characterized by inhibition at the 'preferred' disparity. The half-width at half-height of this peak (or trough) was measured along the two cardinal axes. The orientation of the long axis defined the orientation of the response surface, where zero indicates that the long axis lies horizontally.

Several analyses were performed to examine the possibility that the horizontal orientation of disparity responses was the result of variability in the animals' horizontal vergence angle. All of these suggested that the phenomenon was not attributable to eye movements. For example, the s.d. of vergence (horizontal and vertical) was calculated for each set of trials used. The geometric mean ratio (0.91) was not significantly different from unity (t -test on log ratios). Note also that the broad distribution of preferred horizontal disparities (Fig. 4) cannot readily be explained on the basis of fluctuations in vergence state.

Received 4 March; accepted 10 June 2002; doi:10.1038/nature00909.

- Cumming, B. G. & DeAngelis, G. C. The physiology of stereopsis. *Annu. Rev. Neurosci.* **24**, 203–238 (2001).
- Ohzawa, I. Mechanisms of stereoscopic vision: the disparity energy model. *Curr. Opin. Biol.* **8**, 509–515 (1998).
- Ohzawa, I., DeAngelis, G. C. & Freeman, R. D. Stereoscopic depth discrimination in the visual cortex: Neurons ideally suited as disparity detectors. *Science* **249**, 1037–1041 (1990).
- Mayhew, J. & Longuet-Higgins, H. C. A computational model of binocular depth perception. *Nature* **297**, 376–378 (1982).
- Howard, I. P. & Rogers, B. J. *Binocular Vision and Stereopsis* (Oxford Univ. Press, Oxford, 1995).

- Anzai, A., Ohzawa, I. & Freeman, R. D. Neural mechanisms for processing binocular information II. Complex cells. *J. Neurophysiol.* **82**, 909–924 (1999).
- Barlow, H. B., Blakemore, C. & Pettigrew, J. D. The neural mechanisms of binocular depth discrimination. *J. Physiol. (Lond.)* **193**, 327–342 (1967).
- Joshua, D. E. & Bishop, P. O. Binocular single vision and depth discrimination: Receptive field disparities for central and peripheral vision and binocular interaction of peripheral single units in cat striate cortex. *Exp. Brain Res.* **10**, 389–416 (1970).
- Heydt, R. v. d., Adorjani, C., Anny, P. H. & Baumgartner, G. Disparity sensitivity and receptive field incongruity of units in the cat striate cortex. *Exp. Brain Res.* **31**, 523–545 (1978).
- Maunsell, J. H. R. & van Essen, D. C. Functional properties of neurons in middle temporal visual area of the macaque monkey. II. Binocular interactions and sensitivity to binocular disparity. *J. Neurophysiol.* **49**, 1148–1166 (1983).
- Nieder, A. & Wagner, H. Encoding of both vertical and horizontal disparity in random-dot stereograms by Wulst neurons of awake barn owls. *Vis. Neurosci.* **18**, 541–547 (2001).
- DeAngelis, G. C., Ohzawa, I. & Freeman, R. D. Depth is encoded in the visual cortex by a specialized receptive field structure. *Nature* **352**, 156–159 (1991).
- Hubel, D. H. & Wiesel, T. N. Receptive fields and functional architecture of monkey striate cortex. *J. Physiol. (Lond.)* **195**, 215–243 (1968).
- Cumming, B. G. & Parker, A. J. Responses of primary visual cortical neurons to binocular disparity without depth perception. *Nature* **389**, 280–283 (1997).
- Ohzawa, I., DeAngelis, G. C. & Freeman, R. D. Encoding of binocular disparity by complex cells in the cat's visual cortex. *J. Neurophysiol.* **77**, 2879–2909 (1997).
- Livingstone, M. S. & Tsao, D. Y. Receptive fields of disparity-selective neurons in macaque striate cortex. *Nature Neurosci.* **2**, 825–832 (1999).
- Cumming, B. G. & Parker, A. J. Binocular neurons in V1 of awake monkeys are selective for absolute, not relative, disparity. *J. Neurosci.* **19**, 5602–5618 (1999).
- Cumming, B. G. & Parker, A. J. Local disparity not perceived depth is signalled by binocular neurons in cortical area V1 of the macaque. *J. Neurosci.* **20**, 4758–4767 (2000).
- Prince, S. J. D., Poinson, A. D., Cumming, B. G. & Parker, A. J. The precision of single neuron responses in cortical area V1 during stereoscopic depth judgements. *J. Neurosci.* **20**, 3387–3400 (2000).
- Prince, S. J. D., Cumming, B. G. & Parker, A. J. Range and mechanism of horizontal disparity encoding in macaque V1. *J. Neurophysiol.* **87**, 209–221 (2002).

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.G.C. (e-mail: bgc@lsr.nei.nih.gov).

Wnt-11 activation of a non-canonical Wnt signalling pathway is required for cardiogenesis

Petra Pandur^{*†}, Matthias Läsche^{*}, Leonard M. Eisenberg[‡] & Michael Kühl^{*†}

^{*} Abt Entwicklungsbiologie, Junior group SFB 271, Universität Göttingen, Humboldtallee 23, D-37073 Göttingen, Germany

[†] Abt Biochemie, Universität Ulm, Albert-Einstein-Allee-11, D-89081 Ulm, Germany

[‡] Department of Cell Biology and Anatomy, Medical University of South Carolina, Charleston, South Carolina 29425, USA

Formation of the vertebrate heart requires a complex interplay of several temporally regulated signalling cascades¹. In *Xenopus laevis*, cardiac specification occurs during gastrulation and requires signals from the dorsal lip and underlying endoderm². Among known *Xenopus* Wnt genes, only Wnt-11 shows a spatio-temporal pattern of expression that correlates with cardiac specification, which indicates that Wnt-11 may be involved in heart development^{3,4}. Here we show, through loss- and gain-of-function experiments, that XWnt-11 is required for heart formation in *Xenopus* embryos and is sufficient to induce a contractile phenotype in embryonic explants. Treating the mouse embryonic carcinoma stem cell line P19 with murine Wnt-11 conditioned medium triggers cardiogenesis, which indicates that the function of Wnt-11 in heart development has been conserved in higher vertebrates. XWnt-11 mediates this effect by non-

canonical Wnt signalling, which is independent of β -catenin and involves protein kinase C and Jun amino-terminal kinase. Our results indicate that the cardiac developmental program requires non-canonical Wnt signal transduction.

Wnt proteins signal through several pathways including the non-canonical Wnt/Ca²⁺ and Wnt/Jun N-terminal kinase (JNK) pathways (Wnt-5A class) and the canonical Wnt/ β -catenin pathway (Wnt-1 class)^{5–8}. Studies have suggested that inhibition of Wnt/ β -catenin signalling initiates formation of the vertebrate heart^{9,10}, although the possible involvement of non-canonical Wnt pathways was not analysed. In *Xenopus*, maternal XWnt-11 protein (Wnt-5A class), is enriched in the presumptive dorsal mesoderm⁴ and zygotic expression commences in the dorsal marginal zone (DMZ) where high amounts of XWnt-11 persist during early gastrulation³. Wnt-11 expression in precardiac mesoderm is conserved in mouse and avian embryos^{11–13}.

To investigate whether XWnt-11 is required for cardiac specification in *Xenopus*, we carried out loss-of-function experiments using a dominant-negative (dn) XWnt-11 construct that inhibits XWnt-11 function^{5,14}. Introducing dnXWnt-11 into the presumptive cardiac region strongly decreased subsequent expression of the cardiac markers XNkx2.5 and cardiac troponin I (TnIc; refs 15, 16 and Fig. 1a, b). This effect could be rescued by injecting wild-type XWnt-11 with dnXWnt-11 RNAs (Fig. 1a, b). By contrast, injecting the same amount of dnXWnt-8 RNA, an inhibitor of zygotic Wnt/ β -catenin signalling¹⁴, did not interfere with cardiac gene expression (Fig. 1a, b). Specific inhibition of XWnt-11 signalling in the DMZ by morpholino antisense oligonucleotides (XWnt-11 MO) also interfered with cardiac gene expression (Fig. 1d, e).

Immunoblot analysis with a specific antibody against Wnt-11 (refs 12, 13) verified that injections of XWnt-11 MO significantly reduced the quantities of XWnt-11 protein in DMZ explants, as compared with non-injected control explants (Fig. 1c). The effect of the XWnt-11 MO could be rescued by co-injecting murine Wnt-11 RNA, which is not a target for the antisense MOs (Fig. 1f). Histological analyses of affected embryos showed that there was a general disorganization of the endo- and myocardium (Fig. 1e). Injecting the same amount of a control MO did not interfere with cardiac gene expression or heart formation (Fig. 1f). Together, the dnXWnt-11 and XWnt-11 MO experiments show that XWnt-11 is required for normal heart development in *Xenopus*.

We examined whether cardiac gene expression could be induced in pluripotent animal cap cells in response to overexpression of XWnt-11. Polymerase chain reaction with reverse transcription (RT-PCR) analyses showed that XWnt-11 alone was sufficient to induce the early cardiac genes GATA-4, GATA-6 and XNkx2.5 (Fig. 2a), and α -myosin heavy chain (MHC α) and TnIc (Fig. 2b), which are characteristic of and specific for terminally differentiated cardiomyocytes. Absence of the panmesodermal marker Xbra and the panendodermal marker Sox17 β suggested that cardiac induction by XWnt-11 is direct and independent of germ-layer specification.

Animal cap tissue treated with activin differentiates into mesodermal and endodermal derivatives¹⁷, as monitored by the expression of Xbra and Sox17 β (Fig. 2a) and several cardiac markers (Fig. 2a, b). If XWnt-11 is essential in cardiac development, then activation of cardiac genes by activin should be sensitive to the introduction of dnXWnt-11. Indeed, injections of dnXWnt-11 RNA inhibited activin-induced expression of GATA-4, GATA-6 and XNkx2.5 (Fig. 2a). XWnt-11 and activin also cooperated in the activation of cardiac genes (Fig. 2a, b). Thus, XWnt-11 induces both early cardiac genes and expression of myocardial structural proteins, showing that XWnt-11 is both required and sufficient for cardiac differentiation.

We investigated whether Wnt/ β -catenin signalling mimics XWnt-11 activity in animal caps. Overexpression of β -catenin did not induce the expression of GATA-4, GATA-6, cardiac actin,

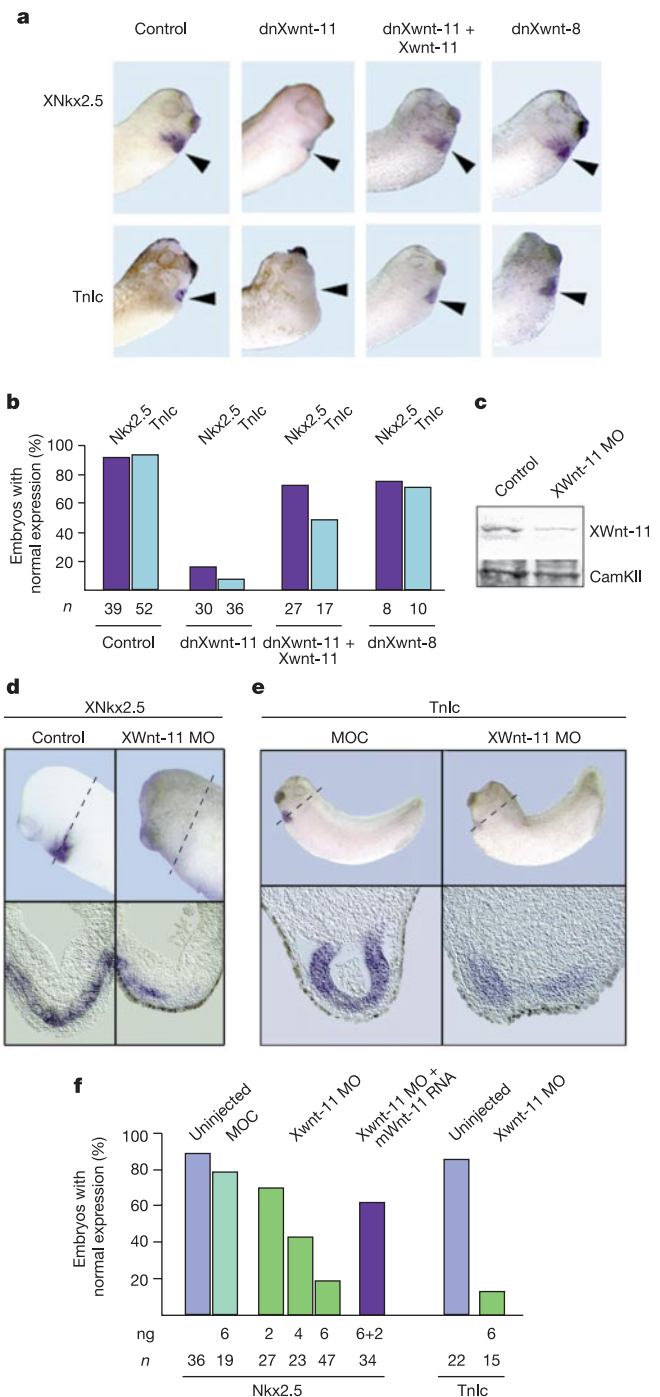


Figure 1 Inhibition of XWnt-11 signalling interferes with heart development. **a**, Embryos were injected in the presumptive cardiac region of the DMZ at the four-cell stage as indicated and analysed for expression of the cardiac marker genes XNkx2.5 and TnIc (arrowheads) by whole-mount *in situ* hybridization at stage 29/30. **b**, Percentage of embryos depicted in **a** showing normal or decreased expression of XNkx2.5 and TnIc, respectively. *n*, number of embryos scored. **c**, Reduction of endogenous XWnt-11 protein in DMZ explants after injections of XWnt-11 MO. CamKII is the loading control. **d**, **e**, Injections of XWnt-11 MO reduce expression of XNkx2.5 and TnIc in contrast to control MO (MOC) injected embryos. **d**, Transverse sections (dashed line) through the heart region of these embryos show the reduction of XNkx2.5 transcripts in the cardiogenic mesoderm. **e**, TnIc is normally expressed exclusively in the myocardium of the linear heart tube as can be seen in the transverse section of a MOC-injected embryo. Injection of XWnt-11 MO but not MOC interferes with primary heart tube formation. **f**, Injected embryos were scored for normal or decreased expression of XNkx2.5 and TnIc as indicated.

MHC α or TnIc (Fig. 2b, c). None of the early cardiac genes was upregulated in a synergistic manner by β -catenin and activin (Fig. 2c). In addition, activin-induced cardiogenesis was not abolished by dnXWnt-8 RNA, as it was with dnXWnt-11 RNA (Fig. 2a, c). As these data were consistent with the observation that dnXWnt-8 does not repress cardiogenesis in whole embryos (Fig. 1a, b), we concluded that XWnt-11 exerts its function independently of β -catenin.

Because XWnt-11 has been implicated to act through non-canonical Wnt-pathways^{5,6}, we tested whether activation of these pathways stimulates cardiogenesis in animal caps. The known mediators of non-canonical Wnt signalling are protein kinase C (PKC)¹⁸, Ca²⁺/calmodulin-dependent kinase II (CamKII)¹⁴ and JNK^{19,20}. In contrast to overexpression of XWnt-11, neither injections of constitutively active CamKII (data not shown) or a dishevelled construct that specifically activates JNK signalling (dsh Δ DIX)^{5,19}, nor incubation with the PKC activator phorbol-12-myristate-13-acetate (PMA) induced cardiac gene expression in animal caps (Fig. 2d). But activation of both PKC and JNK (PMA plus dsh Δ DIX) resulted in robust cardiac gene expression.

We examined whether components of non-canonical Wnt pathways can override dnXWnt-11 inhibition of cardiac gene expression in DMZ explants. We first verified that dnXWnt-11 expression in DMZ explants downregulated cardiac gene transcription (Fig. 2e). We then tested whether inhibitors of PKC or CamKII can phenocopy the effect of dnXWnt-11. Whereas treatment of uninjected

DMZ explants with a PKC inhibitor (bisindolylmaleimide I; BIM) mimicked the effect of dnXWnt-11 on cardiac gene expression (Fig. 2e), specific inhibition of CamKII by KN-93 had no effect. In conjunction with the animal cap experimentation (Fig. 2d), these data implicate the involvement of PKC but not CamKII in XWnt-11-induced cardiogenesis. Consistent with these findings, the expression of cardiac genes in DMZ explants could be rescued from dnXWnt-11 inhibition by treatment with PMA (Fig. 2e). To determine whether JNK activation can compensate for the inhibitory effects of dnXWnt-11, we injected dsh Δ DIX RNA with dnXWnt-11 into the dorsolateral region. RT-PCR analyses of these DMZ explants showed that cardiac gene expression was restored to normal amounts (Fig. 2e). Thus, a block in XWnt-11 signalling can be overcome by activation of JNK. The requirement of JNK signalling for cardiogenic differentiation was also shown by injections of a kinase-dead JNK mutant²¹, JNK-KR, which decreased cardiac gene expression markedly (Fig. 2e).

Because these data showed that JNK was involved in the pathway downstream from XWnt-11, we tested whether XWnt-11 activates JNK-1 by measuring the extent of JNK-1 phosphorylation²². XWnt-11 overexpression in animal halves increased JNK-1 phosphorylation without altering the total amount of JNK-1 present, which indicated that JNK-1 was activated by XWnt-11 (Fig. 2f). Activation of JNK-1 required PKC activity, because the stimulating effect of XWnt-11 was blocked by BIM. XWnt-8, which acts through β -catenin, failed to activate JNK-1 (Fig. 2f). A decrease in endogen-

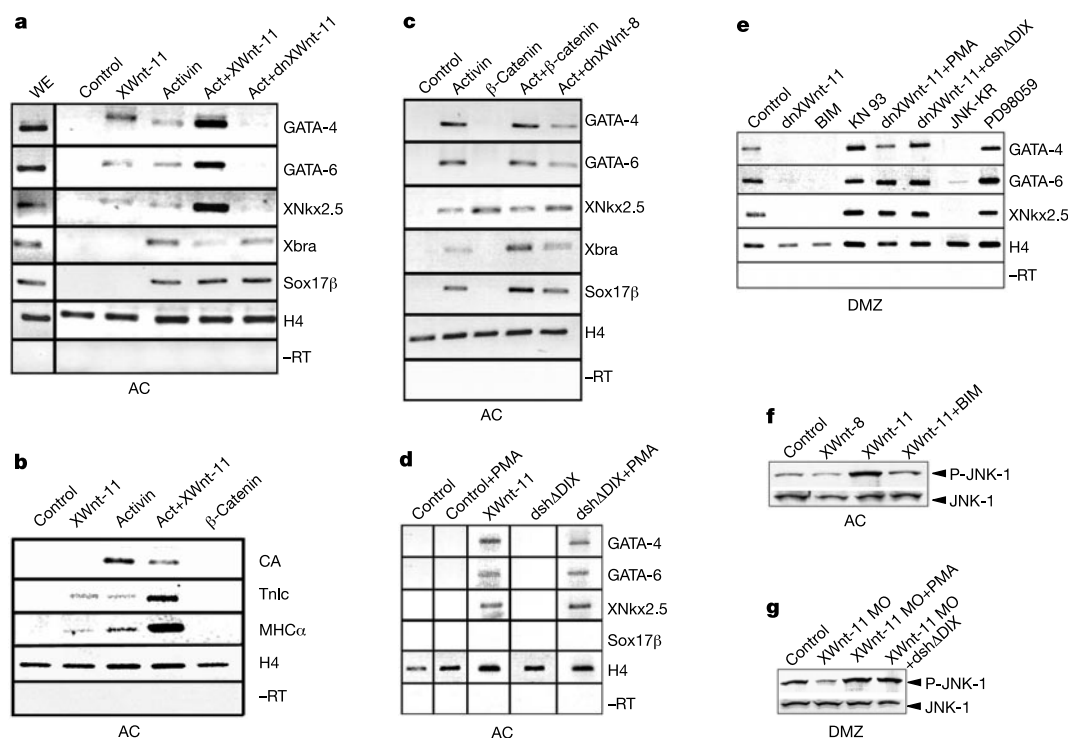


Figure 2 XWnt-11 triggers early steps of cardiogenesis in pluripotent precursor cells by activating JNK. **a**, *Xenopus* embryos were injected as indicated. Marker gene expression in animal caps (AC) was analysed by RT-PCR at stage 19/20. GATA-4, GATA-6 and XNkx2.5, early cardiac marker genes; Xbra, mesodermal marker; Sox17 β , endodermal marker; H4, loading control; -RT, negative control lacking reverse transcriptase; WE, positive control using RNA from whole embryos. **b**, XWnt-11 positively regulates cardiomyocyte-specific marker genes. Animal caps were injected as indicated. Analyses of marker genes were at stage 20 for cardiac actin (CA) and stage 30 for cardiac MHC α and TnIc. **c**, Effect of XWnt-11 on cardiac gene expression in animal caps cannot be mimicked by β -catenin. **d**, Simultaneous activation of PKC and JNK results in the induction of early cardiac marker genes in animal caps comparable to overexpression of

XWnt-11. Animal caps were dissected from uninjected control embryos and from embryos injected with XWnt-11 or dsh Δ DIX RNA and cultured with or without PMA. **e**, Cardiac gene expression in DMZ explants of uninjected and injected embryos in the presence of the indicated agents. PD 98059 specifically inhibits the ERK cascade of MAPK signalling and is a negative control. **f**, Injection of XWnt-11 but not XWnt-8 RNA activates JNK-1 in dissected animal caps, as measured by the phosphorylated form of JNK-1. Activation of JNK-1 by XWnt-11 depends on PKC. The amount of JNK-1 protein remained unchanged in these explants. **g**, Injection of XWnt-11 MO results in a decrease of JNK-1 activity in DMZ explants as measured by JNK-1 phosphorylation. Either treatment with PMA or injection of dsh Δ DIX RNA reversed the effect of XWnt-11 MO in these explants. The amount of JNK-1 protein did not change.

ous phosphorylated JNK-1 in response to XWnt-11 MO injections further supported the regulation of JNK by XWnt-11 (Fig. 2g). Both PMA treatment and co-injection with dsh Δ DIX RNA rescued the effect of XWnt-11 MO on JNK-1 phosphorylation (Fig. 2g), which places PKC upstream of JNK in non-canonical Wnt signalling. Thus, our data indicate that JNK activation is a necessary component of Wnt-11 cardiogenic activity.

We analysed whether the formation of contractile tissue requires XWnt-11 expression. By itself, activin triggers animal caps to develop beating tissue (Fig. 3a and Supplementary Information); however, injection of dnXWnt-11 prevented the formation of activin-induced contractile tissue (Fig. 3a). We found that activin significantly enhanced expression of both XWnt-11 and its receptor XFz-7 (ref. 7) in animal caps (Fig. 3b), further supporting our data that XWnt-11 is required for cardiogenesis.

We also tested whether overexpression of XWnt-11 would induce a contractile phenotype in ventral marginal zone (VMZ) explants¹⁰. XWnt-11 was sufficient to convert non-cardiogenic VMZ explants to contractile tissue (Fig. 3c and Supplementary Information). The potency of the cardiac-inducing activity of XWnt-11 was similar to that of dickkopf-1 (dkk-1; ref. 10 and Fig. 3c). Because dkk-1 has been characterized primarily as a Wnt inhibitor, its previously described cardiogenic activity was assumed to be due to the specific inhibition of Wnt/ β -catenin signalling¹⁰. As Wnt-11 also inhibits Wnt/ β -catenin signalling^{23,24}, we determined whether β -catenin inhibition was responsible for heart formation in VMZ explants.

We inhibited β -catenin-mediated transcriptional activity by expressing a dominant-negative LEF-1 (LEF- Δ HMG)²⁵ in VMZ explants, which competes with all endogenous TCF/LEF (T-cell and lymphoid-enhancer) factors for cytoplasmic β -catenin. This direct approach for inhibiting β -catenin signalling did not induce a contractile phenotype in VMZ explants (Fig. 3c), although identical dnLEF-1 concentrations ventralized *Xenopus* embryos after dorsal injection. Thus, inhibition of β -catenin signalling is not sufficient for heart formation in VMZ explants.

Wnt-11 not only inhibits β -catenin signalling, but also activates JNK. We therefore reasoned that JNK activity might be required for the cardiogenic activity of dkk-1. Injecting JNK-KR and dkk-1 RNAs reduced the heart inducing capacity of dkk-1 (Fig. 3c), which suggested that dkk-1 might not only inhibit Wnt/ β -catenin signalling but also affect Wnt/JNK signalling. Indeed, injections of dkk-1 RNA into animal caps significantly activated JNK-1. In addition, crescent and GSK-3—two other Wnt inhibitors that trigger cardiogenesis¹⁰—also activated JNK, whereas LEF Δ HMG did not (Fig. 3d). In contrast to analyses of non-cardiogenic control tissue, histological analyses of explants that became contractile in response to either dkk-1 or XWnt-11 showed that there were large tubular structures in the contractile areas (Fig. 3e). These structures were reminiscent of the primary heart tube in the developing embryo¹⁰. These data clearly indicate that induction of heart tissue not only requires low Wnt/ β -catenin signalling activity¹⁰, but also high Wnt/JNK activity (Fig. 3f). Given the localized expression of both dkk-1

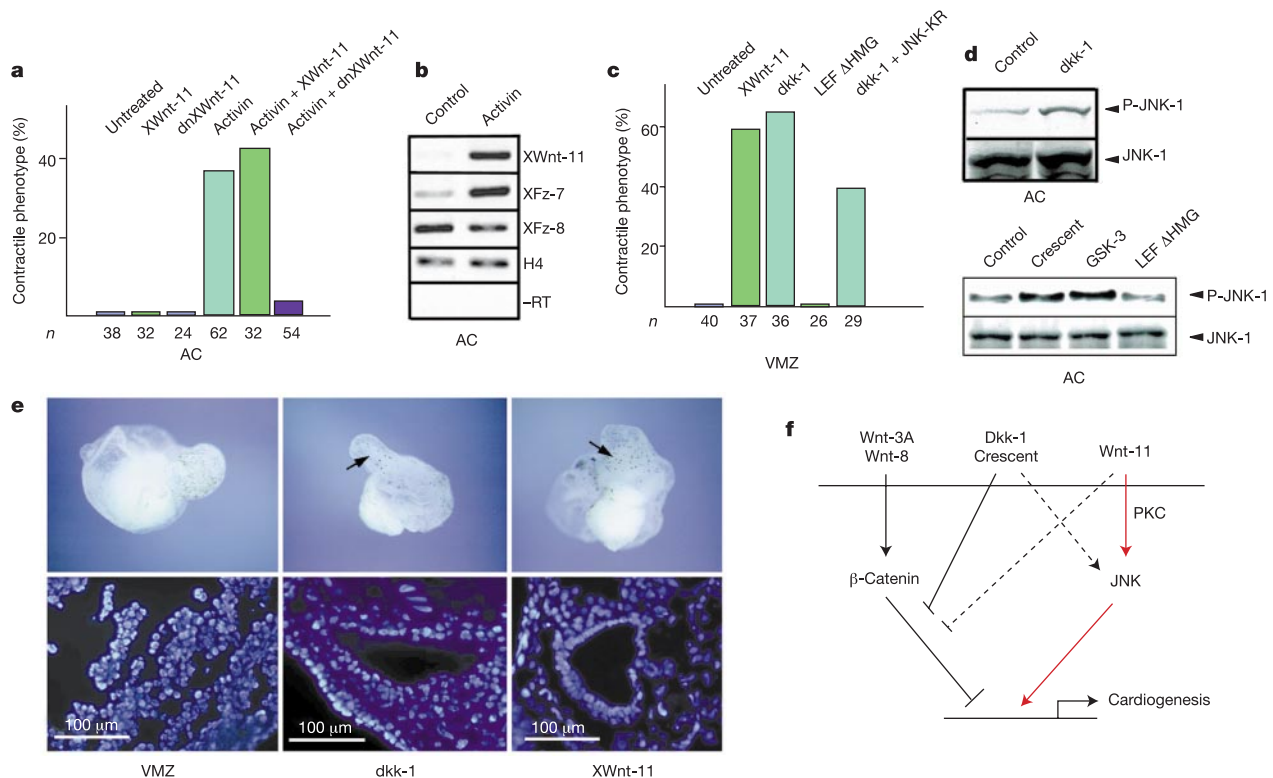


Figure 3 XWnt-11 and formation of contractile tissue in embryonic explants. **a**, Treatment of animal cap explants (AC) with activin (100 ng ml⁻¹) causes formation of beating tissue at stage 42. Injection of dnXWnt-11 blocks formation of activin-induced contractile tissue. *n*, number of animal cap explants scored. **b**, Activin induces expression of XWnt-11 and its receptor XFz-7 in animal cap explants. H4, loading control; -RT, negative control lacking reverse transcriptase. **c**, XWnt-11 induces formation of contractile tissue in VMZ explants at stage 42. Indicated RNAs were injected into the ventrolateral marginal zone at the four-cell stage. *n*, number of VMZ explants scored. **d**, Wnt inhibitors activate JNK in animal cap explants. dkk-1, crescent²⁹, GSK-3 and LEF Δ HMG RNAs were injected, and activation of JNK-1 was measured by phosphorylation of JNK-1. **e**, XWnt-11 induces

formation of tubular structures in VMZ explants that are not detected in untreated explants. Arrows indicate contracting areas of explants. Nuclei were stained with DAPI. Scale bar, 100 μ m. **f**, Model of the interactions of different Wnt pathways in cardiogenesis. *In vivo*, Wnt-3A and Wnt-8 inhibit cardiogenesis through stabilization of β -catenin^{9,10}. Wnt-11 promotes cardiogenesis by activating JNK. The Wnt inhibitors dkk-1 and crescent regulate cardiogenesis positively by blocking β -catenin signalling, which releases the inhibitory effect of Wnt-3A and Wnt-8. In overexpression experiments (dashed lines), dkk-1 and crescent can also activate JNK, whereas Wnt-11 also inhibits β -catenin signalling²³. Thus, these factors promote cardiogenesis by reducing β -catenin signalling activity and increasing JNK signalling activity.

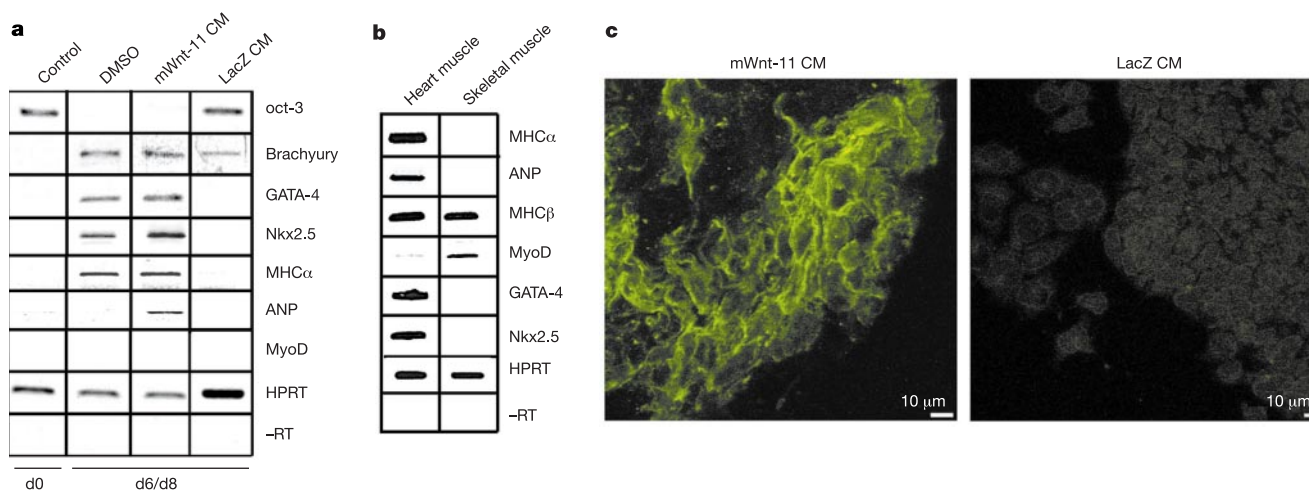


Figure 4 Wnt-11 conditioned medium (CM) triggers cardiomyocyte formation in mouse P19 carcinoma cells. **a**, Murine P19 embryonic carcinoma stem cells express the stem cell marker oct-3. Treatment of P19 cells with DMSO or mWnt-11 CM activates a cardiogenic program as indicated by expression of cardiomyocyte-specific marker genes. Treatment of P19 cells with LacZ CM served as a control. Weak expression of brachyury was also induced by the aggregation of cells in the absence of DMSO. Day 6/8 after the

initial treatment is shown. **b**, MHC α and atrial natriuretic peptide (ANP) are expressed in heart muscle but not skeletal muscle, whereas MyoD is expressed in skeletal muscle but not heart muscle. **c**, mWnt-11 CM but not LacZ CM induced expression of the sarcomeric MHC as analysed by immunofluorescence using the monoclonal antibody MF20. Scale bar, 10 μ m.

and XWnt-11 in or adjacent to the primary heart fields, it is likely that these conditions are provided in this area of the *Xenopus* embryo.

We used the mouse embryonic carcinoma cell line P19 to investigate conservation of the cardiogenic function of Wnt-11 across species. P19 cells undergo cardiomyocyte differentiation in response to dimethyl sulphoxide (DMSO; ref. 26 and Fig. 4a). In response to murine Wnt-11 (mWnt-11) conditioned medium, expression of the stem cell marker oct-3 was abolished and expression of GATA-4 and Nkx2.5 was activated in P19 cells. In addition, expression of the cardiomyocyte markers cardiac MHC α and atrial natriuretic peptide was induced in P19 cells treated with mWnt-11 conditioned medium (Fig. 4a). In contrast, P19 cells cultured in control LacZ conditioned medium showed none of these phenotypic changes (Fig. 4a). The specificity of the primers used in the RT-PCR analyses for cardiac and skeletal muscle marker genes was verified using RNA preparations of adult mouse tissue (Fig. 4b). Cardiomyocytes could also be visualized in cellular aggregates treated with mWnt-11, but not lacZ conditioned medium, by immunofluorescence with an antibody¹⁷ against sarcomeric MHC (Fig. 4c). Thus, as observed in *Xenopus* embryos, mWnt-11 initiates a cardiogenic program in mouse pluripotent precursor cells.

Our results using *X. laevis* and mouse P19 cells clearly show that signalling cascades activated by Wnt-11 are crucial for initiating a cardiac developmental program. Our data also show that both PKC and JNK are essential for the cardiogenic activity of Wnt-11 and provide further evidence on the primary importance of non-canonical Wnt signalling pathways in understanding Wnt functional activity. □

Methods

RNA transcription, microinjections and explant cultures

Capped RNAs were transcribed *in vitro* using the Message Machine Kit (Ambion). For the animal cap assay, embryos were injected at the two-cell stage and caps were dissected at stage 8. The explants were cultured in 1 $\times$ modified Barth solution²⁴ (MBSH) until uninjected control embryos reached stage 19/20. We injected the following amounts of RNA: 1 ng of dnXWnt-11, XWnt-11, dnXWnt-8 or dnLEF-1; 1 pg of activin; 100 pg of β -catenin; 1.5 ng of dsh Δ DIX; 3 ng of JNK-KR; 500 pg of XWnt-8, GSK-3; 800 pg of dkk-1; 900 pg of crescent. JNK was initially isolated from rat²¹. The efficiency of RNAs used for negative controls was tested by different functional assays. Dominant-negative XWnt-8 inhibited XWnt-8-induced axis duplication; dorsal injection of dnLEF-1 RNA ventralized the *Xenopus* embryo. For the analyses of gene expression in DMZ explants, RNAs were injected into the dorsolateral marginal zone at the four-cell stage. DMZ explants were

dissected at stage 10.5 as described² and cultured in 0.5 $\times$ MBSH until sibling embryos reached stage 19/20. For activating or inhibiting PKC, the DMZ explants were cultured in 0.5 $\times$ MBSH containing either PMA or BIM. To inhibit CamKII and MEK1, we added either KN93 or PD 98059. We used concentrations recommended by the manufacturer (Calbiochem). For ectopic induction of beating tissue in VMZ explants, RNAs were injected into the ventrolateral marginal zone at the four-cell stage. VMZ explants were dissected at stage 10.5 and cultured in 0.5 $\times$ MBSH until sibling embryos reached stage 42 (ref. 10).

RNA isolation and RT-PCR analyses

Total RNA was isolated from the explanted tissues using a Purescript RNA Isolation Kit (Biozym). First strand complementary DNAs were synthesized according to the Invitrogen protocol using Superscript II reverse transcriptase. We prepared the PCR reaction mixtures using the PCR Core Kit (Epicentre) and carried out PCR under standard conditions. Sequences of the oligonucleotide primers are available from the authors on request. Amplified products were analysed on 1.2% agarose gels.

In situ hybridization and histology

Whole-mount *in situ* hybridization was carried out by a standard protocol. Digoxigenin-labelled antisense riboprobes of Nkx2.5 (ref. 15) and cardiac troponin (TnI)¹⁶ were synthesized as described. Vibratome sections were cut at 30 μ m thickness and mounted on coverglasses with 90% glycerol, 10% 0.1 M Tris-HCl (pH 7.4). For visualizing nuclei in VMZ explants, cryosections of 14 μ m were cut and mounted with a solution containing 4',6-diamidino-2-phenylindole dihydrochloride (DAPI).

XWnt-11 morpholino antisense oligonucleotide

The antisense oligonucleotide used was a 25-nucleotide morpholino oligo (Gene Tools LLC) with the base composition 5'-CCAGTGACGGGTCGGAGCCATTGGT-3'. A standard morpholino oligo with the base composition 5'-CCTCTTACCTCAGTTAC AATTATA-3' was used as a control.

Western blot and immunoprecipitation

Western blotting and immunoprecipitation studies were done as described¹⁴. Polyclonal and monoclonal antibodies against JNK-1 were purchased from Santa Cruz. We analysed JNK activity at stage 14. The antibody against Wnt-11 was newly produced and tested for specificity as described^{12,13}.

Cell culture

We obtained P19 embryonic carcinoma cells from Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ) (ACC 316). The cells were cultured and induced to differentiate as described²⁸. Stably transfected Wnt-11 and LacZ NIH 3T3 cells were a kind gift from A. Kispert (MH Hannover). To obtain Wnt-11 and LacZ conditioned media, we replaced the culture medium when cells had grown to 50% confluence and collected it after 3 days. The conditioned medium was centrifuged and stored at -20 $^{\circ}$ C. We first characterized the mWnt-11 conditioned medium using two well-known assays for measuring Wnt signalling activity and specificity^{14,18}. As expected, Wnt-11 conditioned medium could activate CamKII in animal caps within 10 min but failed to induce siamois expression (data not shown). The MF20 antibody used for immunofluorescence assays was purchased from the Developmental Studies Hybridoma Bank (University of Iowa). Sequences of the oligonucleotide primers used for RT-PCR analyses are available from the

authors on request.

Received 8 March; accepted 27 May 2002; doi:10.1038/nature00921.

1. Sucov, H. M. Molecular insights into cardiac development. *Annu. Rev. Physiol.* **60**, 287–308 (1998).
2. Nascone, N. & Mercola, M. An inductive role for the endoderm in *Xenopus* cardiogenesis. *Development* **121**, 515–523 (1995).
3. Ku, M. & Melton, D. A. *XWnt-11*: a maternally expressed *Xenopus* wnt gene. *Development* **119**, 1161–1173 (1993).
4. Schroeder, K. E., Condic, M. L., Eisenberg, L. M. & Yost, H. Y. Spatially regulated translation in embryos: asymmetric expression of maternal Wnt-11 along the dorsal-ventral axis in *Xenopus*. *Dev. Biol.* **214**, 288–297 (1999).
5. Tada, M. & Smith, J. C. *XWnt11* is a target of *Xenopus* Brachyury: regulation of gastrulation movements via Dishevelled, but not through the canonical Wnt pathway. *Development* **127**, 2227–2238 (2000).
6. Miller, J. R., Hocking, A. M., Brown, J. D. & Moon, R. T. Mechanism and function of signal transduction by the Wnt/beta-catenin and Wnt/Ca²⁺ pathways. *Oncogene* **18**, 7860–7872 (1999).
7. Djiane, A., Riou, J.-F., Umbhauer, M., Boucaut, J.-C. & Shi, D.-L. Role of frizzled 7 in the regulation of convergent extension movements during gastrulation in *Xenopus laevis*. *Development* **127**, 3091–3100 (2000).
8. Kühl, M., Sheldahl, L. C., Park, M., Miller, J. R. & Moon, R. T. The Wnt/Ca²⁺ pathway: a new vertebrate Wnt signaling pathway takes shape. *Trends Genet.* **16**, 279–283 (2000).
9. Marvin, M. J., Di Roco, G., Gardiner, A., Bush, S. M. & Lassar, A. B. Inhibition of Wnt activity induces heart formation from posterior mesoderm. *Genes Dev.* **15**, 316–327 (2001).
10. Schneider, V. A. & Mercola, M. Wnt antagonism initiates cardiogenesis in *Xenopus laevis*. *Genes Dev.* **15**, 304–315 (2001).
11. Kispert, A., Vainio, S., Shen, L., Rowitch, D. H. & McMahon, A. P. Proteoglycans are required for maintenance of *Wnt-11* expression in the ureter tips. *Development* **122**, 3627–3637 (1996).
12. Eisenberg, C. A., Gourdie, R. G. & Eisenberg, L. M. *Wnt-11* is expressed in early avian mesoderm and required for the differentiation of the quail mesoderm cell line QCE-6. *Development* **124**, 525–536 (1997).
13. Eisenberg, C. A. & Eisenberg, L. M. WNT11 promotes cardiac tissue formation of early mesoderm. *Dev. Dyn.* **216**, 45–58 (1999).
14. Kühl, M., Sheldahl, L. C., Malbon, C. C. & Moon, R. T. Ca²⁺/Calmodulin-dependent protein kinase II is stimulated by Wnt and Frizzled homologs and promotes ventral cell fates in *Xenopus*. *J. Biol. Chem.* **275**, 12701–12711 (2000).
15. Tonissen, K. F., Drysdale, T. A., Lints, T. J., Harvey, R. P. & Krieg, P. A. XNkx-2.5, a *Xenopus* gene related to Nkx-2.5 and tinman: evidence for a conserved role in cardiac development. *Dev. Biol.* **162**, 325–328 (1994).
16. Drysdale, T. A., Tonissen, K. F., Patterson, K. D., Crawford, M. J. & Krieg, P. A. Cardiac troponin I is a heart-specific marker in the *Xenopus* embryo: expression during abnormal heart morphogenesis. *Dev. Biol.* **165**, 432–441 (1994).
17. Grunz, H. Change in the differentiation pattern of *Xenopus laevis* ectoderm by variation of the incubation time and concentration of vegetalizing factor. *Roux's Arch. Dev. Biol.* **192**, 130–137 (1983).
18. Sheldahl, L. C., Park, M., Malbon, C. C. & Moon, R. T. Protein kinase C is differentially stimulated by Wnt and Frizzled homologs in a G-protein-dependent manner. *Curr. Biol.* **9**, 695–698 (1999).
19. Boutros, M. & Mlodzik, M. Dishevelled: at crossroads of divergent intracellular signaling pathways. *Mech. Dev.* **83**, 27–37 (1999).
20. Yamanaka, H. et al. JNK functions in the non-canonical Wnt pathway to regulate convergent extension movements in vertebrates. *EMBO Rep.* **3**, 69–75 (2002).
21. Nishitoh, H. et al. ASK1 is essential for JNK/SAPK activation by TRAF2. *Mol. Cell.* **2**, 389–395 (1998).
22. De Windt, L. J., Lim, H. W., Haq, S., Force, T. & Molkentin, J. D. Calcineurin promotes protein kinase C and c-Jun NH₂-terminal kinase activation in the heart. *J. Biol. Chem.* **275**, 13571–13579 (2000).
23. Torres, M. A. et al. Activities of the wnt-1 class of secreted factors are antagonized by the wnt-5 A class and by a dominant negative cadherin in early *Xenopus* embryo. *J. Cell Biol.* **133**, 1123–1137 (1996).
24. Kühl, M. et al. Antagonistic regulation of convergent extension movements by Wnt/beta-catenin and Wnt/Ca²⁺ signalling. *Mech. Dev.* **106**, 61–76 (2001).
25. Behrens, J. et al. Functional interaction of beta-catenin and the architectural transcription factor LEF-1. *Nature* **382**, 638–642 (1996).
26. Grepin, C., Nemer, G. & Nemer, M. Enhanced cardiogenesis in embryonic stem cells overexpressing the GATA-4 transcription factor. *Development* **124**, 2387–2395 (1997).
27. Bader, D., Masaki, T. & Fischman, D. A. Immunohistochemical analysis of myosin heavy chain during avian myogenesis *in vivo* and *in vitro*. *J. Cell Biol.* **95**, 763–770 (1982).
28. Rudnicki, M. A. & McBurney, M. W. in *Teratocarcinomas and Embryonic Stem Cells. A Practical Approach* (ed. Robertson, E. J.) 19–49 (IRL, Oxford, 1987).
29. Pera, E. M. & DeRobertis, E. M. A direct screen for secreted proteins in *Xenopus* embryos identifies distinct activities for the Wnt antagonists crescent and Frzb-1. *Mech. Dev.* **96**, 183–195 (2000).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements

We thank R. T. Moon, A. McMahon, P. Krieg, K. Matsumoto, E. Pera, D. Kimelman, W. Birchmeier and A. Kispert for providing cDNA clones; A. Kispert for providing the Wnt-transfected NIH3T3 cells; D. Gradl for help with confocal imaging analyses; and T. Hollemann for help with video capturing.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for material should be addressed to M.K. (e-mail: wnt_signaling@web.de).

A component of the ARC/Mediator complex required for TGFβ/Nodal signalling

Yoichi Kato*, Raymond Habas*†, Yu Katsuyama*†, Anders M. Näär‡ & Xi He*

* Division of Neuroscience, Children's Hospital, Department of Neurology, Harvard Medical School, Boston, Massachusetts 02115, USA

‡ Department of Cell Biology, Harvard Medical School, Massachusetts General Hospital Cancer Center, Charlestown, Massachusetts 02129, USA

The transforming growth factor β (TGFβ) family of cytokines, including Nodal, Activin and bone morphogenetic protein (BMP), have essential roles in development and tumorigenesis^{1,2}. TGFβ molecules activate the Smad family of signal transducers, which form complexes with specific DNA-binding proteins to regulate gene expression^{1,2}. Two discrete Smad-dependent signalling pathways have been identified: TGFβ, Activin and Nodal signal via the Smad2 (or Smad3)–Smad4 complex, whereas BMP signals via the Smad1–Smad4 complex^{1,2}. How distinct Smad complexes regulate specific gene expression is not fully understood. Here we show that ARC105, a component of the activator-recruited co-factor (ARC)³ complex or the metazoan Mediator complex, is essential for TGFβ/Activin/Nodal/Smad2/3 signal transduction. Expression of ARC105 stimulates Activin/Nodal/Smad2 signalling in *Xenopus laevis* embryos, inducing axis duplication and mesoderm differentiation, and enhances TGFβ response in human cells. Depletion of ARC105 inhibits TGFβ/Activin/Nodal/Smad2/3 signalling and *Xenopus* axis formation, but not BMP/Smad1 signalling. ARC105 protein binds to Smad2/3–Smad4 in response to TGFβ and is recruited to Activin/Nodal-responsive promoters in chromatin in a Smad2-dependent fashion. Thus ARC105 is a specific and key ARC/Mediator component linking TGFβ/Activin/Nodal/Smad2/3 signalling to transcriptional activation.

The human ARC³ and related or identical metazoan Mediator^{4,5}, TRAP⁶, DRIP⁷, SMCC⁸, CRSP⁹ and PC2¹⁰ have emerged as central players in transcription regulation^{11,12}. These complexes consist of 20–25 polypeptides, and can associate with various transcription factors and stimulate transcription *in vitro*^{3–12}. Some metazoan Mediator subunits are homologous to those in yeast Mediator^{11,12} and may participate in the interaction between Mediator and RNA polymerase II. Other metazoan-specific Mediator components may serve as adapters that link specific transcription activators and/or signalling pathways to basal transcription apparatus^{11,12}.

In a functional screen for complementary DNAs that induce neural gene expression, we isolated the *Xenopus* homologue of ARC105 (*XARC105*), an ARC/Mediator component of unknown function^{3,13,14}. *XARC105* is 68% and 29% identical to human and *Drosophila* ARC105 (*dARC105*), respectively, each harbouring a central glutamine-rich domain. *XARC105* messenger RNA is expressed maternally and widely throughout embryogenesis, with a higher neural tissue expression (Supplementary Information Fig. 1). Ventral injection of *XARC105* RNA, but not RNA for several other ARC subunits (ARC34, 36, 70, 77, 130, 150 and 205; ref. 3), unexpectedly duplicated the dorsal axis (98%, *n* = 55), which contained muscle and neural tissues but lacked a head and notochord (Fig. 1a). Animal pole explants expressing XARC105 protein

† Present addresses: Laboratory of Molecular Genetics, The National Institute of Child Health and Human Development, NIH, Bethesda, Maryland 20892, USA (R.H.); Department of Developmental and Cell Biology, University of California, Irvine, California 92697, USA (Y.K.).

exhibited strong morphogenetic elongation (Fig. 1b). These effects are indistinguishable from those resulting from ectopic Nodal/Activin/Smad2 signalling^{2,15}. Indeed, *XARC105* and *Activin* acted synergistically at low doses in axis duplication (not shown).

Nodal/Activin/Smad2 signalling causes mesendoderm induction accompanied by secondary neural induction and suppression of ventral and epidermal differentiation^{2,15}. Indeed, *XARC105* induced mesendodermal genes in a dose-dependent manner in animal pole explants: it induced at lower doses *brachyury* (*Xbra*, early pan-mesoderm) and *chordin* (dorsal mesoderm), and, at a higher dose, *mixer* (endoderm) (Fig. 1c). *dARC105* behaved similarly and induced *Xbra* and *chordin* (Fig. 1d), indicating functional conservation. *XARC105* also induced other early and late mesendodermal and neural markers, but not *Xnr3*, which is induced by Wnt/ β -catenin signalling¹⁵, and suppressed *vent-1*, *vent-2* and other ventral/epidermal markers induced by BMP/Smad1 signalling¹⁵ (Fig. 1e and Supplementary Information Fig. 2). Therefore, *XARC105* phenotypically copies Nodal/Activin/Smad2 signalling fully.

We examined where in the Nodal/Activin/Smad2 pathway *XARC105* functions. A dominant negative Activin receptor mutant, $\Delta 1XAR1$ (ref. 16), completely blocked *Activin* but not *XARC105* induction of *Xbra*, *chordin* and *mixer* (Fig. 2a), suggesting that *XARC105* acts downstream of the Nodal/Activin receptor. The effect of *XARC105* expression was blocked by a dominant negative Smad2 (*DNSmad2*)—but not a dominant negative Smad1 (*DNSmad1*)¹⁷—and by *FAST-1SID*¹⁸, a dominant negative *FAST-1* mutant (Fig. 2a). *FAST-1* is a DNA-binding protein that associates with and recruits the Smad2–Smad4 complex^{2,18}. These data suggest that *XARC105* function depends on the endogenous Smad2 and *FAST-1*, but not on Smad1.

To examine whether *XARC105* is required for Nodal/Activin/Smad2 signalling, we reduced the endogenous *XARC105* protein

level with an antisense Morpholino oligonucleotide (MO), which blocked *XARC105* translation initiation but did not affect protein levels for actin, tubulin (Fig. 2e) or β -catenin (not shown). In animal pole explants from embryos injected with the *XARC105* MO, induction of *Xbra*, *chordin* and *mixer* by *Activin*, *Xnr1* (*Xenopus* Nodal-related 1) or Smad2 was specifically inhibited (Fig. 2b). This inhibition was rescued by injection of *dARC105* RNA (Fig. 2b), whose translation is resistant to the *XARC105* MO interference. A control MO had no effect on *XARC105* protein level or Activin/Nodal/Smad2 signalling (Fig. 2b, e). *XARC105* MO affected neither BMP/Smad1 induction of *vent-1* and *vent-2* nor *Xwnt-8*/ β -catenin induction of *Xnr3* and *siamoi*s (Fig. 2c, d). In the whole embryo, *XARC105* MO, but not the control MO, inhibited the expression of *chordin* and *gsc*, but not *Xnr3* in the dorsal region, and the inhibition was rescued by *dARC105*; by contrast, *vent-1*/*vent-2* expression induced by BMP/Smad1 signalling in the ventral region was not affected (Fig. 2f and Supplementary Information Fig. 2). Most embryos injected dorsally with *XARC105* MO failed to gastrulate properly (Fig. 2f), because Nodal signalling is required for gastrulation^{2,15}. The few embryos that survived exhibited severe dorso-anterior deficiency (not shown). Therefore, *XARC105* is specifically required for Nodal/Activin/Smad2, but not BMP/Smad1 or Wnt/ β -catenin signalling.

We also examined whether *ARC105* is required for TGF β /Activin signalling in human 293T cells. We reconstituted Activin response with transfected *FAST-1* and *ActRIB**, an activated Activin receptor, and found that an ARE (Activin-responsive element)-driven promoter¹⁹ was stimulated dose dependently by *XARC105* (Fig. 3a). Similarly, TGF β response assayed with 3TP-luc or SBE (Smad3-binding element)-luc was also enhanced by *XARC105* (Fig. 3b and Supplementary Information Fig. 3). By contrast, *XARC105* exhibited a small but consistent suppression of a BMP-responsive

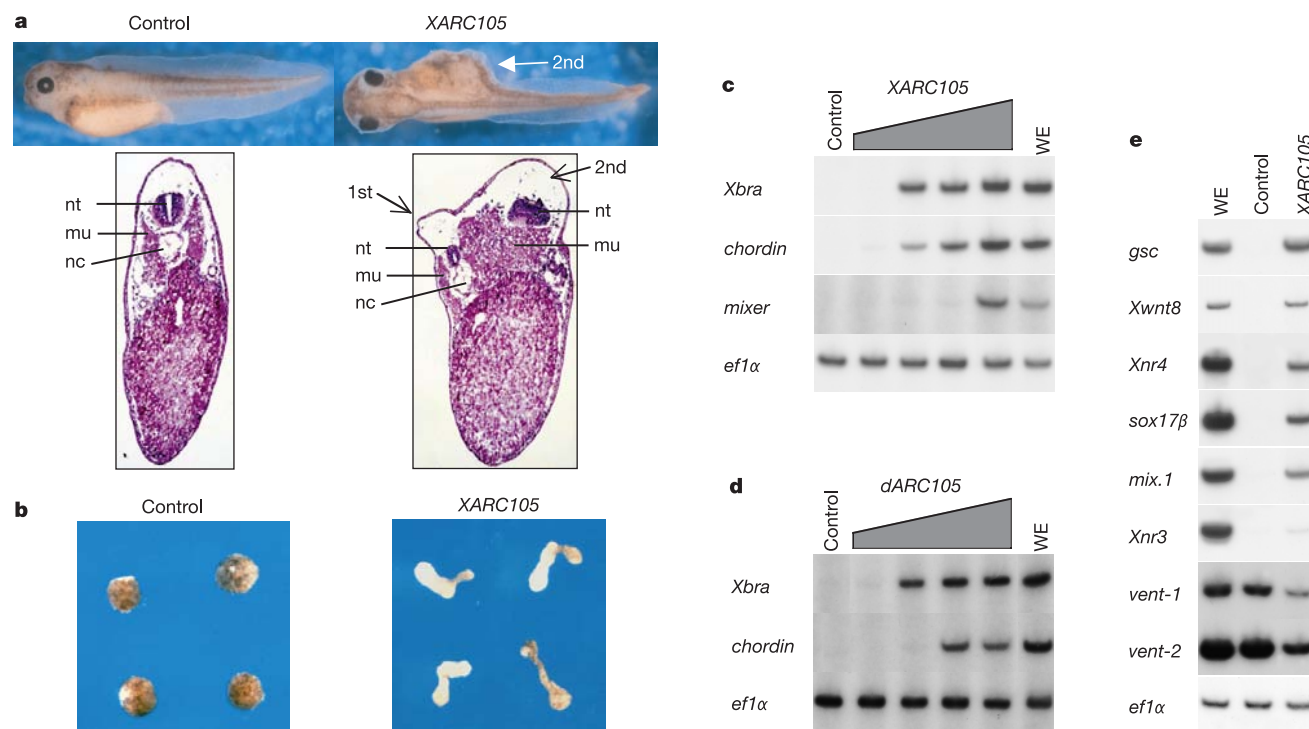


Figure 1 *XARC105* induces axis duplication, morphogenetic elongation and mesendoderm formation. **a**, *XARC105* RNA (1 ng) injected ventrally induced a secondary axis. *XARC105* protein expression was five times the endogenous level at this RNA dose (not shown). Top, stage 35 embryos; bottom, cross-section histology. 1st, primary axis; 2nd, induced axis; mu, muscle; nc, notochord; nt, neural tissues. **b–e**, Animal pole explants. **b**, *XARC105* RNA (1 ng) induced explant elongation at stage 20. **c–e**, RT-PCR

for gene expression at stage 10. Controls are uninjected. WE, whole embryos. *ef1 α* , loading control for RT-PCR. **c**, **d**, RNA for *XARC105* (0.1, 0.5, 1, 2 ng) (**c**) and *dARC105* (0.5, 1, 2.5, 5 ng) (**d**) induced mesoderm/endoderm markers in a dose-dependent manner. **e**, *XARC105* RNA (2 ng) induced markers for early mesoderm (*goosecoid* (*gsc*) and *Xwnt-8*) and endoderm (*sox17 β* , *mix.1*, *Xnr4*), but not the dorsal marker *Xnr3*, and suppressed ventral markers *vent-1* and *vent-2*.

promoter, Xvent-luc¹⁷ (Fig. 3c). We depleted the endogenous ARC105 protein in 293T cells by double-strand RNA-mediated interference (siRNA; Fig. 3d). TGF β and Activin responses assayed using 3TP-luc, SBE-luc and ARE-luc were diminished in ARC105-depleted cells, but rescued by *dARC105* (Fig. 3d, e, and Supplementary Information Fig. 3), whereas the BMP response assayed by Xvent-luc was not changed (Fig. 3f). Thus ARC105 is required for TGF β /Activin but not BMP signalling in human cells.

The mutual functional dependence of *XARC105* and *Smad2* prompted us to investigate potential interactions between *XARC105* and *Smad* proteins. *XARC105* bound to *Smad2* and *Smad3* upon TGF β 1 or Activin stimulation in human 293T cells or *Xenopus* embryos (Fig. 4a, b, and Supplementary Information Fig. 4). *XARC105* also bound to *Smad4*, and this interaction was significantly enhanced by TGF β 1 (Fig. 4a) or Activin (Supplementary Information Fig. 4). *XARC105* did not associate with *Smad1* with or without BMP7 stimulation, and its interaction with *Smad4* was abolished by BMP7 (Fig. 4a) as *Smad4* was recruited into the *Smad1*–*Smad4* complex in response to BMP7 (refs 1 and 2, and data

not shown). ARC130, another subunit of the ARC/Mediator complex³, did not associate with *Smad2* or *Smad4*, with or without TGF β 1 (Supplementary Information Fig. 4). Thus *XARC105* associates with *Smad2/3*–*Smad4* in response to TGF β /Activin, but not with *Smad1*–*Smad4* upon BMP signalling.

Smad proteins contain an amino-terminal MH1 domain, a linker and a carboxy-terminal MH2 domain^{1,2}. *Smad2C* and *Smad4C*²⁰, which harbour the MH2 domain, but not *Smad2NL* and *Smad4NL*²⁰, which contain the MH1 domain and the linker, interacted with *XARC105* (Fig. 4c). Thus the MH2 domain, which carries transcriptional activation function^{1,2}, mediates association between *XARC105* with *Smad2* and *Smad4*.

To examine whether the endogenous *Smad2/3*–*Smad4* complex and the ARC/Mediator complex interact, we immunoprecipitated *Smad2/3* from the nuclear extract of 293T cells untreated or treated with TGF β 1, which stimulated nuclear translocation of *Smad2/3* and *Smad4*, as well as association of *Smad2/3* and *Smad4* (refs 1, 2) (Fig. 5a). The ARC/Mediator complex was nuclear with or without TGF β 1, but became associated with *Smad2/3*–*Smad4* in TGF β 1-

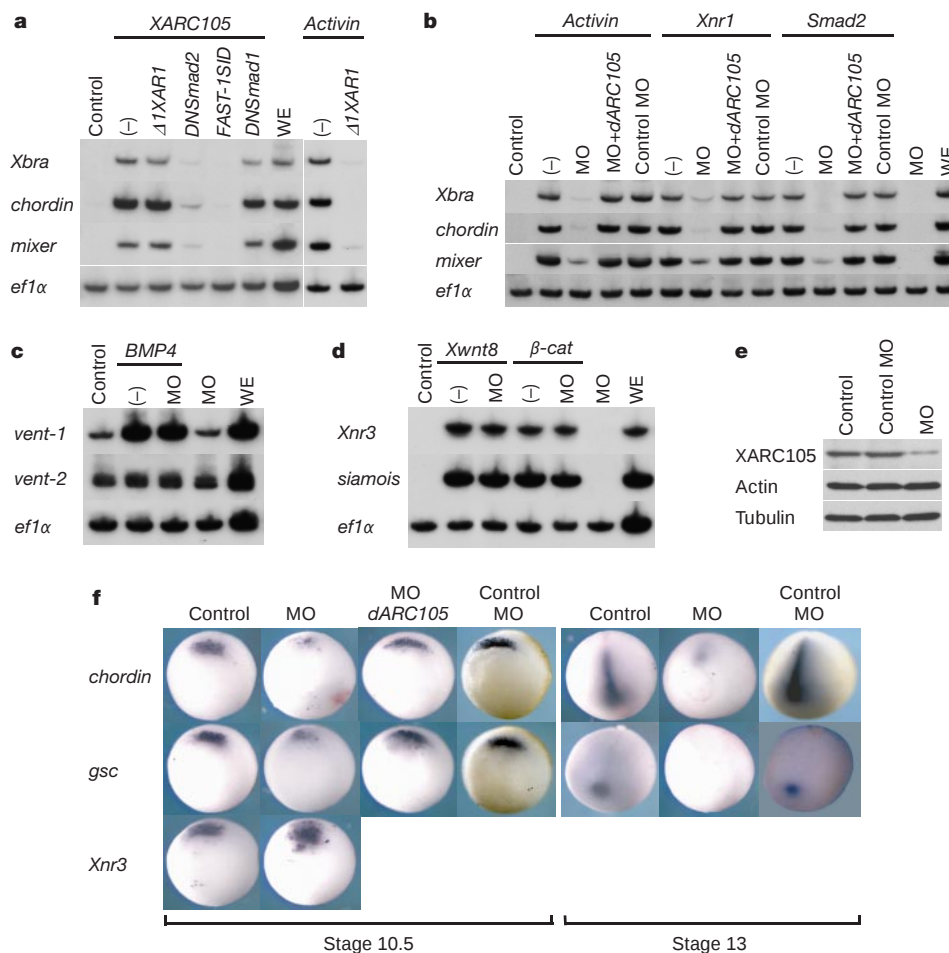


Figure 2 *XARC105* is required for Nodal/Activin/Smad2 signalling. **a–d**, RT-PCR in animal pole explants at stage 10. **a**, *XARC105* induction of *Xbra*, *chordin* and *mixer* was inhibited by dominant negative *Smad2* (DNSmad2) and FAST-1SID, but not by dominant negative *Smad1* (DNSmad1) or Δ 1XAR1. DNSmad2 and DNSmad1 are truncation mutants that lack the C-terminal part of the MH2 domain¹⁷. We injected 2 ng of RNA for each except *Activin* (10 pg). **b**, *XARC105* Morpholino oligonucleotide (MO), but not the control MO, inhibited induction of *Xbra*, *chordin* and *mixer* by injected RNA for *Activin*, *Xnr1* (200 pg) or *Smad2* (2 ng). *dARC105* RNA (2 ng) rescued the inhibition. **c**, **d**, *XARC105* MO did not affect *vent-1* or *vent-2* expression induced by endogenous or BMP4 RNA (2 ng), nor *Xnr3* or *siamois* expression induced by *Xwnt-8* (10 pg) or β -catenin

RNA (200 pg). **e**, Immunoblotting of embryos at stage 10. *XARC105* MO, but not control MO, reduced the level of *XARC105* protein but not of actin or tubulin. **f**, Whole-mount *in situ* hybridization. *XARC105* MO, but not control MO, inhibited organizer-specific expression of *chordin* and *gsc*, but not *Xnr3* at stage 10.5, and inhibited *chordin* and *gsc* expression in anterior mesoderm and blocked gastrulation at stage 13. *XARC105* MO inhibition was rescued by *dARC105*. Stage 10.5 embryos are shown in vegetal view (dorsal at the top); stage 13 embryos are shown in dorsal view (anterior at the bottom), except for embryos injected with the *XARC105* MO, which are shown in vegetal and posterior view, illustrating residual *chordin* expression near the open blastopore.

stimulated cells, as demonstrated by the co-precipitation of ARC105, ARC100, ARC95 and ARC33/Med6 with Smad2/3–Smad4 (Fig. 5a). In contrast, nuclear laminA and laminC were not co-precipitated (Fig. 5a). Thus, TGF β stimulates the formation and nuclear translocation of the Smad2/3–Smad4 complex, which then associates with ARC/Mediator to activate transcription.

We were intrigued to find that the Smad1–Smad4 complex, which entered nuclei upon BMP signalling (Fig. 5a), also associated with ARC/Mediator (but not laminA/C) upon BMP stimulation (Fig. 5a). This was not unexpected, because it is unlikely that the Smad1–Smad4 complex activates transcription through a completely different mechanism. However, there is a key distinction between the Smad1–Smad4–ARC complex and the Smad2/3–Smad4–ARC complex. Reducing the level of ARC105 protein by siRNA decreased the amount of not only ARC105 but also other ARC/Mediator components associated with Smad2/3–Smad4 (Fig. 5a), implying that these components associate with Smad2/3–Smad4 via ARC105. By contrast, ARC105 siRNA reduced the amount of ARC105 but not of other ARC/Mediator subunits associated with Smad1–Smad4 (Fig. 5a). These data suggest that the Smad1–Smad4 complex engages the ARC/Mediator complex through an ARC subunit distinct from ARC105, consistent with ARC105 binding to neither Smad1 nor Smad4 upon BMP signalling (Fig. 4a).

Because Nodal/Activin induction of gene expression requires XARC105 (Fig. 2), we examined whether XARC105 is recruited to

Nodal/Activin responsive enhancers (NREs) *in vivo* by use of chromatin immunoprecipitation (ChIP). XARC105 associated with, upon Activin stimulation, two distinct NREs from the *mix.2* and *gsc* promoters in chromatin in *Xenopus* embryos, and this association was blocked by *DNSmad2* (Fig. 5b). Thus, XARC105 is recruited to Nodal/Activin responsive promoters by Smad2 during embryogenesis. In the same ChIP experiment, XARC105 was not crosslinked with the BMP-responsive *Xvent-2* promoter upon BMP (or Activin) stimulation (Fig. 5b). Thus in early embryos, Smad1–Smad4 may recruit an ARC/Mediator lacking XARC105. Or, because XARC105 does not bind and is not in close proximity to Smad1–Smad4 and DNA, crosslinking of XARC105 to the Xvent-2

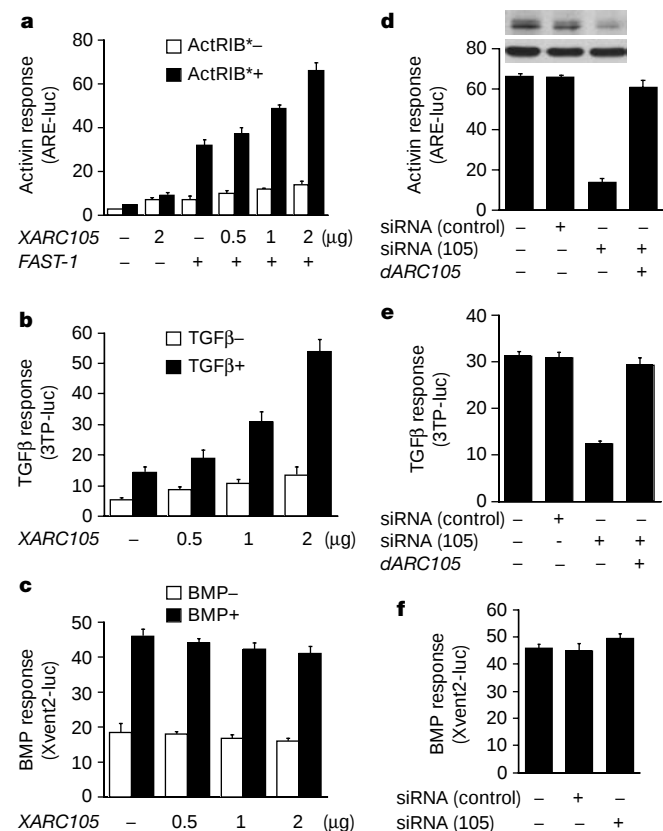


Figure 3 ARC105 is required for TGF β /Activin but not BMP signalling in 293T cells. **a–c**, XARC105 stimulated Activin response from ARE-luc by transfected cDNAs for FAST-1 (200 ng) and ActRIB* (1 μ g), and TGF β 1 response from 3TP-luc, but weakly suppressed BMP7 response from Xvent2-luc. **d**, **e**, ARC105 siRNA, but not a control siRNA, reduced protein level levels of ARC105 (**d**, inset, top) but not tubulin (bottom), and inhibited ARE-luc expression induced by co-transfected FAST-1 and ActRIB* cDNAs and 3TP-luc expression induced by TGF β 1. Inhibition was rescued by dARC105 cDNA (2 μ g). **f**, ARC105 siRNA had no effect on Xvent2-luc expression induced by BMP7.

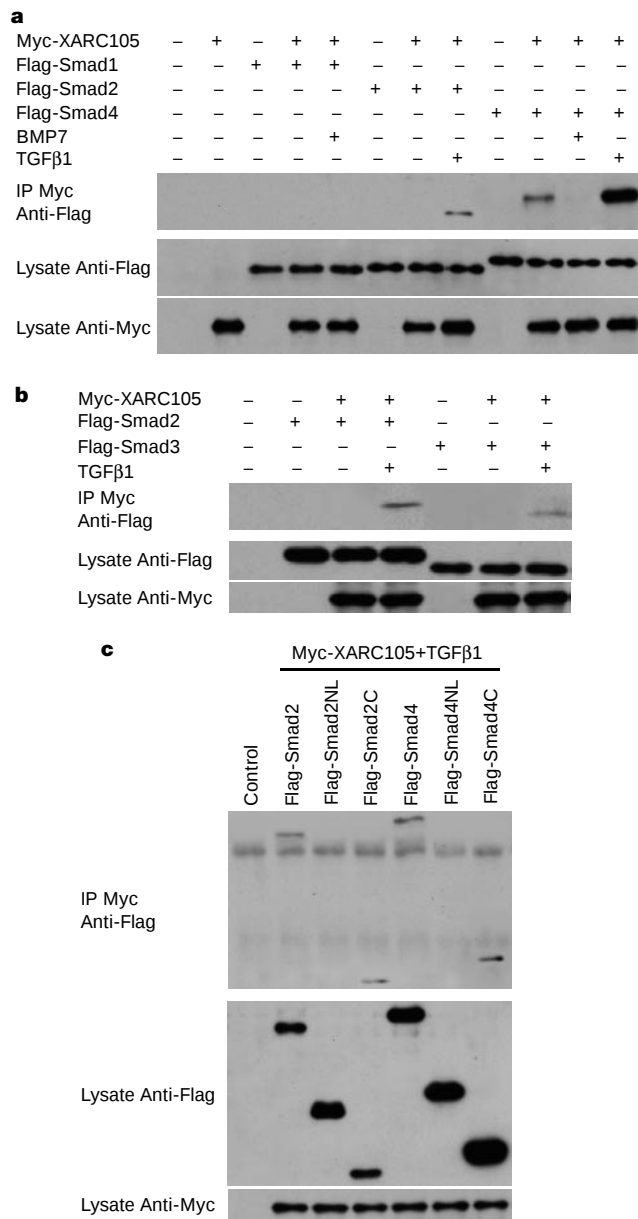


Figure 4 ARC105 protein associates with Smad2/3 and Smad4 in response to TGF β in 293T cells (western blots). **a**, TGF β 1 induced XARC105–Smad2 binding and enhanced XARC105–Smad4 binding, which was abolished by BMP7. XARC105 did not bind Smad1 with or without BMP7. XARC105 binding to Smad4 but not Smad2 when overexpressed in the absence of TGF β 1 may be explained by the existence of proteins such as SARA¹. **b**, TGF β 1 induced XARC105–Smad3 binding. Note that Smad3 was expressed at a lower level than Smad2. **c**, XARC105 interacted with the MH2 domain (Smad2C and Smad4C), but not with the MH1 domain plus the linker (Smad2NL and Smad4NL).

promoter through intermediate ARC subunits becomes inefficient.

We provide compelling evidence that ARC105 specifically mediates TGF β /Activin/Nodal/Smad2/3 signalling. Overexpression of *ARC105* activates all aspects of Nodal/Smad2 signalling in *Xenopus* embryos, and enhances TGF β /Activin response in human cells, probably owing to the ability of overexpressed ARC105 to promote the association between the ARC/Mediator and the basal or activated Smad2/3–Smad4 complexes. Depletion of ARC105 inhibits TGF β /Activin/Nodal/Smad2/3 signalling without affecting BMP/Smad1 or Wnt/ β -catenin signalling. This functional specificity is paralleled by ARC105–Smad binding specificity. ARC105 associates with Smad2/3 in response to TGF β /Activin, but does not associate with Smad1 with or without BMP. Additionally, binding of ARC105 to Smad4 is enhanced by TGF β /Activin but is suppressed by BMP. These results suggest that ARC105 is a specific adapter that couples Smad2/3–Smad4 to the ARC/Mediator complex. BMP signalling stimulates Smad1–Smad4 association with the ARC/Mediator complex, probably through a distinct ARC/Mediator subunit.

The ARC/Mediator and related complexes bind to and act as co-activators *in vitro* for several transactivators, including SREBP, NF- κ B, nuclear receptors, p53 and VP16^{11,12}. TRAP220/DRIP205/ARC205 binds to nuclear receptor transactivation domains and is required for certain nuclear receptor functions, whereas TRAP80/ARC77 recognizes both VP16 and p53 transactivation motifs *in vitro*^{11,12}. Genetic studies in invertebrates have revealed distinct and

specific functions for ARC/Mediator subunits^{21–24}, but the underlying molecular basis remains unknown. ARC105 provides an example for a specific signal transduction function of an ARC/Mediator component. It is probable that ARC/Mediator, which interacts with RNA polymerase II, governs TGF β signalling together with other Smad-binding co-activators and co-repressors, such as CBP/p300, Ski/SnoN, MSG1 and SNIP1^{1,2}, Swift²⁵ and SMIF²⁶. The specific function of ARC105 in signalling by Smad2/3–Smad4, which are human tumour suppressors¹, suggests that ARC105 itself is a tumour suppressor. □

Methods

cDNA library construction and screening

A maternal cDNA library was made using the SuperScript Plasmid System (Life Technologies). cDNAs larger than 1 kilobase were ligated into the *Sall* and *NotI* sites in the CS105 vector (ref. 27). The library was divided into 500 pools, with 100 independent cDNAs per pool. We injected 5 ng of RNAs synthesized *in vitro* from each pool into the animal pole of ultraviolet-treated embryos at the 2-cell stage. Positive pools were identified by upregulation of *zic-r1* (an early neural marker) expression at stage 14 using polymerase chain reaction with reverse transcription (RT–PCR). Sib-selections were performed as described²⁷ on positive pools until single cDNAs were isolated. From 12,500 cDNAs screened, *XARC105* and *noggin* were each isolated twice. Four other positive clones are new and yet to be characterized.

Plasmid construction

Myc-tagged *XARC105* and *dARC105* were subcloned into CS105. Human ARC105 was cloned into the CS2 + vector.

Experiments with *Xenopus* embryos

All embryological manipulations, including injection, animal pole, dorsal and ventral marginal zone explants, *in situ* hybridization and RT–PCR, were performed as described²⁸.

Morpholino oligonucleotides

The *XARC105* MO, 5'-CAACACCTGCTTCTCCTCATTGCTTC-3', which is complementary to the translation initiation region, was synthesized by Gene Tools and injected at 40 ng per embryo into the animal pole at the 2-cell stage (for animal pole explants) or dorsally or ventrally at the 4-cell stage (for dorsal or ventral marginal zone explants, or embryo phenotypes). An MO with a random sequence of the same length was used as a negative control. The endogenous XARC105 level was examined by immunoblotting with antisera that were generated against the carboxyl 265-amino-acid residues of XARC105. The antisera specifically recognize both *Xenopus* and human ARC105 (not shown).

Luciferase assay

We co-transfected 0.5 μ g of indicated luciferase reporter and 10 ng of Renilla luciferase plasmid (as an internal control) into each well of a six-well plate with indicated expression plasmids by the standard calcium phosphate method. The total amount of DNA transfected was equalized by control vectors without inserts. Transfected cells were treated with TGF β 1 (5 ng ml⁻¹) or BMP7 (100 ng ml⁻¹) for 16 h, and lysed for luciferase measurement using the Dual Luciferase System (Promega).

siRNA

siRNA for human ARC105 in 293T cells was performed as described²⁹. An *ARC105* siRNA oligonucleotide, 5'-UCCUGGCCAACAUCGCGCUC(TT)-3', and its complementary RNA strand also harbouring TT at the 3' end were synthesized by Integrated DNA Technologies, annealed *in vitro*, and co-transfected into 293T cells (1 μ g per well, six-well plates) with indicated plasmids by Lipofectamine (Life Technologies). An siRNA pair with identical length and design for an unrelated human gene was used as a negative control. Cell extracts for luciferase assays and immunodetection of ARC105, actin and tubulin were made 45 h after siRNA transfection.

Immunoprecipitation

Human 293T cells co-transfected with *XARC105* together with *Smad* expression plasmids (2 μ g each) were treated with TGF β 1 (10 ng ml⁻¹) or BMP7 (100 ng ml⁻¹) for 1 h, and lysed with lysis buffer (150 mM NaCl, 1% NP-40, 5 mM EDTA at pH 8.0, 10 mM Tris-HCl at pH 7.5, Complete protease inhibitor (Roche), 30 mM okadaic acid, and 50 mM calyculin A). Lysates were immunoprecipitated with anti-Myc polyclonal antisera (Santa Cruz) and GammaBind G-Sepharose (Amersham Pharmacia Biotech). Precipitates were separated by SDS–polyacrylamide gel electrophoresis (PAGE) and immunoblotted with anti-Flag (M2, Sigma), or anti-Myc (9E10) monoclonal antibodies.

We prepared 293T cell nuclear extracts as described³ and incubated them for 4 h at 4 °C with antibodies specific for Smad1 or for Smad2/3 (Upstate Biotechnology) and GammaBind G-Sepharose in 0.1 M KCl buffer (20 mM Tris-HCl at pH 8.0, 20% glycerol, 100 mM KCl, 1.5 mM MgCl₂, 0.2 mM EDTA, 1 mM dithiothreitol, Complete protease inhibitor (Roche), 30 mM okadaic acid and 50 mM calyculin A). The beads were washed seven times in 0.2 M NaCl, 0.25% NP-40 and 0.1 M KCl. Precipitates were separated by SDS–PAGE and immunoblotted with antibodies against ARC subunits or laminA/C or Smad4 (Santa Cruz Biotechnology).

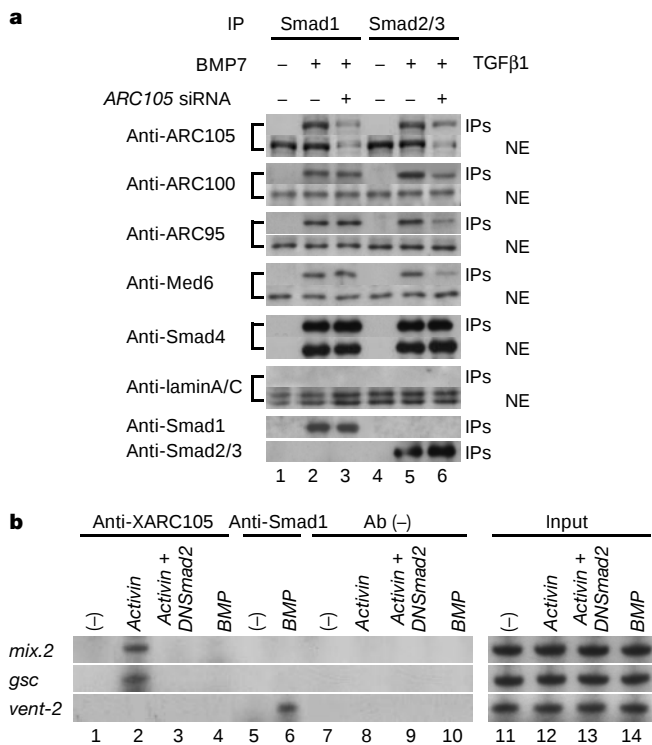


Figure 5 Association between endogenous ARC/Mediator and Smad2/3–Smad4 complex and recruitment of ARC105 to chromatin via Smad2. **a**, Endogenous Smad–ARC complexes in the nuclear extract (NE) of 293T cells treated with TGF β 1 (lanes 4–6) or BMP7 (lanes 1–3). Immunoprecipitation (IP) was done with Smad2/3- or Smad1-specific antibodies. Immunoprecipitates (IPs) and starting NE were immunoblotted with the indicated antibodies. *ARC105* siRNA reduced the protein level of ARC105 but not of other ARC subunits or Smads. **b**, ChIP in *Xenopus* animal pole explants. The promoter region of *mix.2* and *gsc* were co-immunoprecipitated with the endogenous *XARC105* upon Activin signalling, and this co-immunoprecipitation was abolished by *DNSmad2*. The *Xvent-2* promoter was co-immunoprecipitated with *Smad1* upon BMP stimulation, but not with *XARC105* with or without BMP or Activin. Ab (-), immunoprecipitation without antisera. Input, chromatin DNA in cross-linked DNA–protein mix before immunoprecipitation. RNA for *Activin* (1 ng), *BMP4* (2 ng) and *DNSmad2* (2 ng) was injected.

Chromatin immunoprecipitation

Nuclei were isolated from stage 10 *Xenopus* embryos, and ChIP was performed as described³⁰ using a kit from Upstate Biotechnology. We used 20 µl of XARC105 antisera. After PCR, primers flanking these Nodal/Activin- or BMP-responsive enhancers/promoters were used: 5'-GCCTTCACTACAGACACCTT-3' and 5'-AGGCTGCGAAGCCATGACTT-3' (*mix.2*; 38 cycles); 5'-CAGCTCAATGAATGGGTTA-3' and 5'-GAAAAATCGCAGACTCTCCC-3' (*gsc*; 35 cycles); 5'-TAAATTTGACCACTACTGGG-3' and 5'-TTACACATTCTGCCTCTCAC-3' (*vent-2*; 35 cycles).

Received 4 April; accepted 9 July 2002; doi:10.1038/nature00969.

Published online 24 July 2002.

- Massague, J. How cells read TGF-β signals. *Nature Rev. Mol. Cell Biol.* **1**, 169–178 (2000).
- Whitman, M. Smads and early developmental signalling by the TGFβ superfamily. *Genes Dev.* **12**, 2445–2462 (1998).
- Näär, A. M. *et al.* Composite co-activator ARC mediates chromatin-directed transcriptional activation. *Nature* **398**, 828–832 (1999).
- Jiang, Y. W. *et al.* Mammalian mediator of transcriptional regulation and its possible role as an end-point of signal transduction pathways. *Proc. Natl Acad. Sci. USA* **95**, 8538–8543 (1998).
- Boyer, T. G., Martin, M. E. D., Lees, E., Ricciardi, R. P. & Berk, A. J. Mammalian Srb/Mediator complex is targeted by adenovirus E1A protein. *Nature* **399**, 276–279 (1999).
- Fondell, J. D., Ge, H. & Roeder, R. G. Ligand induction of a transcriptionally active thyroid hormone receptor coactivator complex. *Proc. Natl Acad. Sci. USA* **93**, 8329–8333 (1996).
- Rachez, C. *et al.* Ligand-dependent transcription activation by nuclear receptors requires the DRIP complex. *Nature* **398**, 824–828 (1999).
- Gu, W. *et al.* A novel human SRB/MED-containing cofactor complex, SMCC, involved in transcription regulation. *Mol. Cell* **3**, 97–108 (1999).
- Ryu, S., Zhou, S., Ladurner, A. G. & Tjian, R. The transcriptional cofactor complex CRSP is required for activity of the enhancer-binding protein Sp1. *Nature* **397**, 446–450 (1999).
- Malik, S., Gu, W., Wu, W., Qin, J. & Roeder, R. G. The USA-derived transcriptional coactivator PC2 is a submodule of TRAP/SMCC and acts synergistically with other PCs. *Mol. Cell* **5**, 753–760 (2000).
- Näär, A. M., Lemon, B. D. & Tjian, R. Transcriptional coactivator complexes. *Annu. Rev. Biochem.* **70**, 475–501 (2001).
- Rachez, C. & Freedman, L. P. Mediator complexes and transcription. *Curr. Opin. Cell Biol.* **13**, 274–280 (2001).
- Abraham, S. & Solomon, W. B. A novel glutamine-rich putative transcriptional adaptor protein (TIG-1) preferentially expressed in placental and bone-marrow tissues. *Gene* **255**, 389–400 (2000).
- Berti, L. *et al.* Isolation and characterization of a novel gene from the DiGeorge chromosomal region that encodes for a mediator subunit. *Genomics* **74**, 320–332 (2001).
- Harland, R. & Gerhart, J. Formation and function of Spemann's organizer. *Annu. Rev. Cell Dev. Biol.* **13**, 611–667 (1997).
- Hemmati-Brivanlou, A. & Melton, D. A. A truncated activin receptor inhibits mesoderm induction and formation of axial structures in *Xenopus* embryos. *Nature* **359**, 609–614 (1992).
- Candia, A. F. *et al.* Cellular interpretation of multiple TGF-β signals: intracellular antagonism between activin/BVg1 and BMP-2/4 signalling mediated by Smads. *Development* **124**, 4467–4480 (1997).
- Chen, X. *et al.* Smad4 and FAST-1 in the assembly of activin-responsive factor. *Nature* **389**, 85–89 (1997).
- Yeo, C.-Y., Chen, X. & Whitman, M. The role of FAST-1 and Smads in transcriptional regulation by activin during early *Xenopus* embryogenesis. *J. Biol. Chem.* **274**, 26584–26590 (1999).
- Stroschein, S. L., Wang, W., Zhou, S., Zhou, Q. & Luo, K. Negative feedback regulation of TGF-β signalling by the SnoN oncoprotein. *Science* **286**, 771–774 (1999).
- Singh, N. & Han, M. *sur-2*, a novel gene, functions late in the let-60 ras-mediated signalling pathway during *Caenorhabditis elegans* vulval induction. *Genes Dev.* **9**, 2251–2265 (1995).
- Zhang, H. & Emmons, S. W. A *C. elegans* mediator protein confers regulatory selectivity on lineage-specific expression of a transcription factor gene. *Genes Dev.* **14**, 2161–2172 (2000).
- Boube, M., Faucher, C., Joulia, L., Cribbs, D. L. & Bourbon, H. M. *Drosophila* homologs of transcriptional mediator complex subunits are required for adult cell and segment identity specification. *Genes Dev.* **14**, 2906–2917 (2000).
- Treisman, J. E. *Drosophila* homologue of the transcriptional coactivation complex subunits TRAP240 and TRAP230 are required for identical processes in eye-antennal disc development. *Development* **128**, 603–615 (2001).
- Shimizu, K., Bourillot, P.-V., Nielsen, S. J., Zorn, A. M. & Gurdon, J. B. Swift is a novel BRCT domain coactivator of Smad2 in transforming growth factor β signalling. *Mol. Cell Biol.* **21**, 3901–3912 (2001).
- Bai, R.-Y. *et al.* SMIE, a Smad4-interacting protein that functions as a co-activator in TGFβ signalling. *Nature Cell Biol.* **4**, 181–190 (2002).
- Hsu, D. R. *et al.* The *Xenopus* dorsalizing factor Gremlin identifies a novel family of secreted proteins that antagonize BMP activities. *Mol. Cell* **1**, 673–683 (1998).
- Kato, Y., Shi, Y. & He, X. Neuralization of the *Xenopus* embryo by inhibition of p300/CREB-binding protein function. *J. Neurosci.* **19**, 9364–9373 (1999).
- Elbashir, S. M. *et al.* Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411**, 494–498 (2001).
- Sachs, L. M. & Shi, Y.-B. Target chromatin binding and histone acetylation in vivo by thyroid hormone receptor during amphibian development. *Proc. Natl Acad. Sci. USA* **24**, 13138–13143 (2000).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements

We apologize for our inability to cite many original papers owing to space constraints. We thank Z. Chang, K. Cho, E. DeRobertis, R. Harland, A. Hata, A. Hemmati-Brivanlou, D. Kessler, K. Luo, J. Massague, J.-P. Saint-Jeanet, Y. Shi, W. Solomon, G. Thomson, T. Wang, M. Whitman, Y. Zhang and Y. Etoh for reagents; P. Dikkes for help; Y. Sun and M. Greenberg for comments; and K. Luo, Y. Shi, T. Wang, X. Wang and members of the He lab for discussion. This work is supported by postdoctoral fellowships from the Uehara

Memorial Foundation (Japan) and Charles A. King Trust and the Medical Foundation to Y.K., from US Department of Defense (DOD) to R.H. A.M.N. is supported by the Bertucci Foundation. X.H. is supported by grants from the Rockefeller Brothers Fund, DOD and NIH, and is a Pew Scholar, Klingenstein Fellow, and Keck Foundation Distinguished Young Scholar.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to X.H. (e-mail: xi.he@tch.harvard.edu). The GenBank accession number for XARC105 is AF378334.

Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein

Ann M. Sheehy*, Nathan C. Gaddis*, Jonathan D. Choi† & Michael H. Malim*‡

* Department of Microbiology; and † Division of Human Genetics, The Children's Hospital of Philadelphia; University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA

‡ Department of Infectious Diseases, Guy's, King's and St Thomas' School of Medicine, King's College London, London SE1 9RT, UK

Viruses have developed diverse non-immune strategies to counteract host-mediated mechanisms that confer resistance to infection. The Vif (virion infectivity factor) proteins are encoded by primate immunodeficiency viruses, most notably human immunodeficiency virus-1 (HIV-1). These proteins are potent regulators of virus infection and replication and are consequently essential for pathogenic infections *in vivo*^{1–6}. HIV-1 Vif seems to be required during the late stages of virus production^{3,6} for the suppression of an innate antiviral phenotype that resides in human T lymphocytes^{7,8}. Thus, in the absence of Vif, expression of this phenotype renders progeny virions non-infectious. Here, we describe a unique cellular gene, *CEM15*, whose transient or stable expression in cells that do not normally express *CEM15* recreates this phenotype, but whose antiviral action is overcome by the presence of Vif. Because the Vif:CEM15 regulatory circuit is critical for HIV-1 replication, perturbing the circuit may be a promising target for future HIV/AIDS therapies.

Viruses deficient in *vif* (Δvif) display only their non-infectious

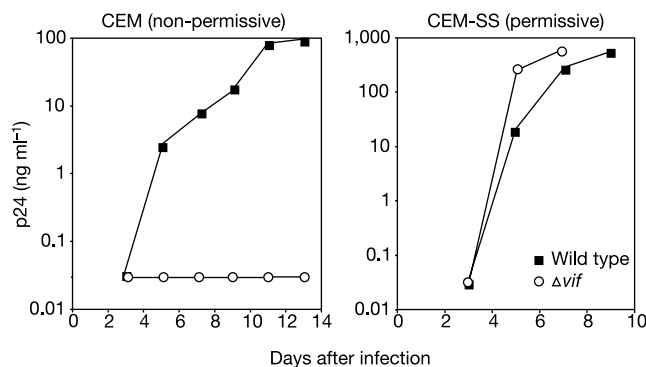


Figure 1 Replication of HIV-1 in non-permissive and permissive T cells. CEM and CEM-SS cells were challenged with normalized stocks of wild-type or Δvif X4-tropic viruses, and virus replication was monitored as the supernatant accumulation of p24^{Gag}.

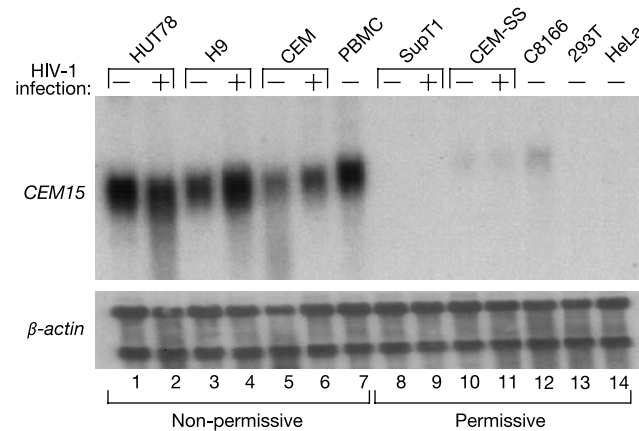


Figure 2 Northern analysis of *CEM15* expression in non-permissive and permissive cells. RNAs extracted from the indicated panel of infected or uninfected cells were resolved by electrophoresis and subjected to northern analysis using a ^{32}P -labelled *CEM15* probe (top

panel). PBMC, peripheral blood mononuclear cells. After autoradiography, the filter was stripped and rehybridized with a β -actin-specific probe to establish equivalent loading (bottom panel).

phenotype when produced by primary human T cells, the principal cell target for HIV-1 *in vivo*, and a limited number of cell lines, including HUT78 and CEM. Such cells are termed non-permissive^{3,6-8}. Conversely, many other cell lines, such as SupT1, CEM-SS and 293T, support the production of fully infectious Δvif virions. These cells are termed permissive. Recent cell-fusion experiments have established that the non-permissive phenotype is dominant over the permissive phenotype in that Δvif virions produced from non-permissive $\times$ permissive heterokaryons are non-infectious^{7,8}. This finding supports a model in which non-permissive cells selectively express an antiviral activity that is suppressed by Vif, and predicts that non-permissive cells express genes whose ectopic expression in permissive cells should phenotypically convert them to non-permissive cells.

To identify candidate genes of this type, we used a complementary DNA subtraction strategy based on polymerase chain reaction (PCR) in which the non-permissive sample served as the 'tester' and the permissive sample was used as the 'driver'. To enrich for clones specific to the non-permissive phenotype, we identified a pair of

genetically related cell lines that exhibit distinct abilities to support Δvif virus replication. CEM-SS is a subclonal isolation of the CEM parental T-cell line^{9,10}, and the genetic relatedness of these two lines was confirmed by sequencing the V(D)J DNA rearrangement junctions at the T-cell receptor- γ locus (data not shown). The non-permissive phenotype of the CEM line was confirmed by its inability to support a spreading infection of HIV-1/ Δvif , whereas the permissive phenotype of CEM-SS cells was verified by the indistinguishable replication profiles of wild-type and Δvif viruses (Fig. 1).

Individual subtracted cDNAs were used as probes in northern analyses of RNAs extracted from a panel of non-permissive and permissive cells. The cDNA corresponding to a gene we have named *CEM15* yielded a particularly marked result (Fig. 2), where a single transcript of about 1.5 kilobases was readily detected in all non-permissive cells tested (lanes 1–7), including peripheral blood mononuclear cells, but minimal (lanes 10–12) or no (lanes 8, 9, 13 and 14) expression was observed in a variety of lymphoid and non-lymphoid permissive cell lines. Notably, expression levels of

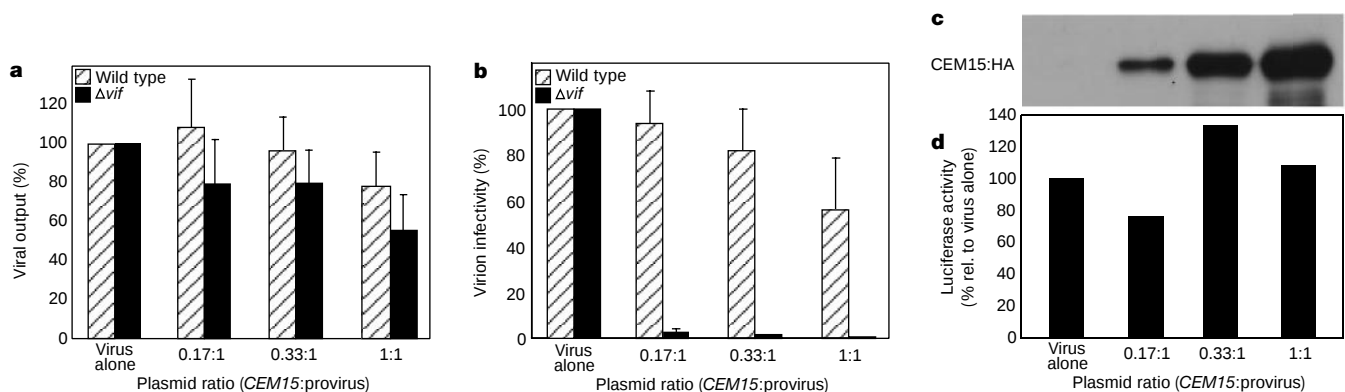


Figure 3 Expression of *CEM15* in permissive cells inhibits the infectivity of HIV-1/ Δvif . We transfected 60-mm-diameter 293T monolayers with cocktails comprising 3 μg provirus (wild type or Δvif), 12 μg pcDNA3.1-based vector (titration of pCEM15:HA with the empty parental vector making up the balance), and 100 ng pLUC. By using 0, 0.5, 1 or 3 μg pCEM15:HA, the indicated ratios of CEM15:provirus were achieved. **a**, Analysis of virus output. Virus-containing supernatants were collected ~24 h after transfection, and virus production quantified as p24^{Gag} by ELISA; the mean $\pm$ s.d. values for eight independent experiments are shown. Identical results were obtained using a CEM15 expression vector that lacked an epitope tag (data not shown). **b**, Single-round virus

infectivity. Viruses from **a** corresponding to 5 ng p24^{Gag} were used in challenges of C8166-CCR5/HIV-CAT cells, and productive infection was measured as the induction of CAT activity. Values are presented as per cent infectivity relative to virus-alone samples (that is, no pCEM15:HA used), and were derived from eight experiments. **c**, Expression of CEM15:HA. The titration of protein expression is confirmed for one series of Δvif transfections by immunoblot analysis of whole-cell lysates using the HA-specific monoclonal antibody 16B12. **d**, Luciferase expression. The cell lysates used for **c** were evaluated for luciferase expression; the values are displayed as per cent activity relative to the virus-alone sample.

CEM15 messenger RNA were not influenced significantly by HIV-1 infection (lanes 2, 4, 6, 9 and 11).

Because the expression profile for *CEM15* correlates with that expected for a gene important for the non-permissive phenotype, we next examined the effects of *CEM15* expression on HIV-1 particle production and infectivity using transiently transfected 293T cells (a permissive cell line). Varying doses of pCEM15:HA (a vector that encodes an epitope-tagged version of the predicted 384-amino-acid open reading frame of *CEM15*) were co-transfected with either wild-type or Δ vif versions of an X4 provirus expression vector and an internal control luciferase-expressing plasmid. At about 24 h, the levels of virus output were assessed by measuring levels of p24^{Gag} protein in the culture supernatants, and were found to be relatively unaffected for either virus by the presence of *CEM15* (Fig. 3a).

These same virus-containing supernatants were then normalized according to p24^{Gag} content and used in single-round challenges of an indicator cell line in which productive HIV-1 infection is registered as the induction of chloramphenicol acetyl transferase (CAT) expression (Fig. 3b). In contrast to the wild-type virus, which was largely unaffected by *CEM15* expression, the Δ vif virus displayed minimal levels of infection at all doses of *CEM15* tested (for example, ~97% reduction in infectivity with a sixfold excess of Δ vif provirus over pCEM15:HA, relative to provirus alone). Whole-cell lysates from one set of Δ vif transfections were also examined by immunoblotting to confirm the titration curve for *CEM15*:HA expression (Fig. 3c), and for luciferase expression to establish that *CEM15* does not affect cell physiology or gene expression in a pleiotropic manner (Fig. 3d). Finally, analogous transfection exper-

iments were also carried out using a matched pair of wild-type and Δ vif R5 proviral clones (see Supplementary Information Fig. S1). The effects of *CEM15* on virus infectivity were essentially the same as for the X4 viruses, thus demonstrating that the inhibition of Δ vif infection by *CEM15* is not strain specific. Taken together, these data reveal that *CEM15* shows all the features expected for a gene that establishes the non-permissive phenotype; that is, its expression in permissive cells inhibits Δ vif virus production qualitatively (inhibition of infectivity), but not quantitatively.

To verify the antiviral function of *CEM15* in an alternative experimental configuration, its effect on HIV-1 replication in T cells was determined. CEM-SS cells were stably transduced with the *CEM15*:HA encoding retroviral vector NG/C15, or with the parental vector NG/neo, and protein expression in the bulk cultures was confirmed by immunoblot analysis of whole-cell lysates (Fig. 4a). Both cell lines were challenged with varying doses of wild-type or Δ vif HIV-1, and the spread of the infection was monitored as the time-dependent accumulation of p24^{Gag} in the culture supernatants (Fig. 4b and Supplementary Information Fig. S2). As noted above for parental CEM-SS cells (Fig. 1), the *neo*-expressing control cells supported efficient replication by both viruses. In contrast, the *CEM15*-expressing cells allowed robust replication by the wild-type virus, but low or no detectable Δ vif virus growth (for example, for the challenges with an intermediate inoculum (Fig. 4b), there was a roughly 300-fold reduction in output of Δ vif virus relative to wild-type virus at day 10). Although we suspect that a lack of adequate *CEM15* expression in a subset of the cells accounts for the residual levels of HIV-1/ Δ vif replication in cultures challenged with higher doses, the possible contribution of other genes to the complete non-permissive phenotype cannot be ruled out¹¹. Nevertheless, these data further demonstrate that ectopic expression of *CEM15* is highly effective at converting cells from the permissive phenotype to the non-permissive phenotype.

HIV-1 virions contain 10–100 molecules of Vif per average particle^{12–14}. To determine whether *CEM15* influences Vif incorporation, and whether *CEM15* itself can be packaged, wild-type and Δ vif virions produced in 293T cells in the presence or absence of *CEM15*:HA were purified and analysed by immunoblotting (Fig. 5). On the basis of this transfection experiment, Vif packaging is not affected by *CEM15* (lanes 1 and 3), and *CEM15* can be incorporated into virions irrespective of Vif expression (lanes 3 and 4).

A database search using *CEM15* identified a gene expressed in haematopoietic cells isolated from myelodysplastic syndrome (MDS) patients (AF182420, *Homo sapiens* phorbolin-like protein

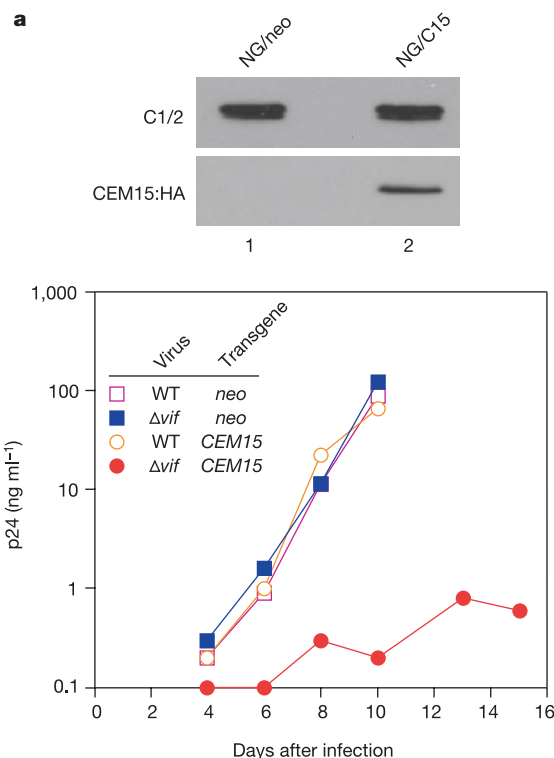


Figure 4 Stable expression of *CEM15* in CEM-SS cells selectively inhibits HIV-1/ Δ vif replication. **a**, Transduction of CEM-SS cells. Cultures were transduced with the NG/neo (negative control) or NG/C15 retroviral vectors, maintained in selective medium, and subjected to immunoblotting using antibodies specific for hnRNP C1/2 (loading control) or the HA epitope tag. **b**, Replication of HIV-1. Control or *CEM15*-expressing cultures were challenged with wild-type or Δ vif virus corresponding to 1 ng p24^{Gag} and monitored as for Fig. 1. WT, wild type.

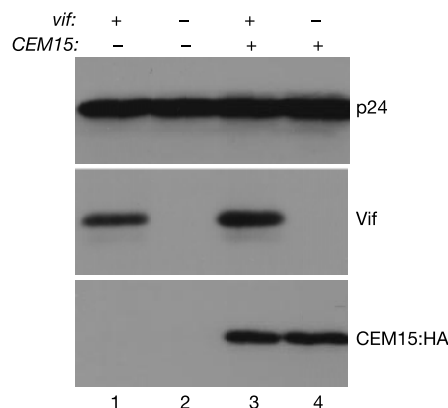


Figure 5 Incorporation of *CEM15* into virions. We transfected 293T monolayers with pIIIB or pIIIB/ Δ vif and pCEM15:HA or empty vector at a 2.5:1 ratio. Viruses were collected at ~24 h, purified with continuous sucrose gradients, formed into pellets by centrifugation, and then analysed by immunoblotting using antibodies specific for p24^{Gag}, Vif or the HA epitope tag. Analysis of p24^{Gag} content (top panel) established equal loading of samples.

MDS019); there is no further description of the gene itself or its function. Analysis of the CEM15 protein sequence (predicted relative molecular mass 46,405) revealed a murine orthologue, but no significant matches in *Saccharomyces cerevisiae*, *Drosophila melanogaster* or *Caenorhabditis elegans*. Further analysis also demonstrated that the amino- and carboxy-terminal portions of human and murine CEM15 possess significant amino-acid similarity to each other, to apobec-1 (the cytidine deaminase that is the catalytic subunit of the mammalian apolipoprotein B mRNA editing enzyme)¹⁵, and to the protein phorbolin-1 (Fig. 6)¹⁶. For instance, the C-terminal region of human CEM15 exhibits about 50% amino-acid identity to apobec-1 and about 70% amino-acid identity to phorbolin-1. Particularly intriguing is the delineation of a zinc-coordinating motif in each of these protein domains (Fig. 6, boxed regions). This zinc-binding domain has previously been identified in cytosine nucleoside/nucleotide deaminases present in organisms ranging from bacteriophages to humans¹⁷, and has been shown to be critical for the catalytic activity of these enzymes^{18,19}. Whether the residues in CEM15 that correspond to those required for deaminase function also operate in rendering Δ vif viruses non-infectious in non-permissive cells remains to be determined.

CEM15 is an efficient and specific inhibitor of HIV-1/ Δ vif infectivity. Because CEM15 is selectively expressed in non-permissive cells (Fig. 2), and its anti-HIV phenotype is readily overcome by the levels of Vif expressed during normal HIV-1 replication (Fig. 4), we assert that the CEM15 protein is the human cell target for Vif function. The sequence similarity between CEM15 and cytosine deaminases is potentially provocative (Fig. 6). Because apobec-1 binds to and edits mRNA¹⁸, and because Vif has been reported not only to bind viral genomic RNA but also to depend on this

interaction for normal incorporation into virions^{20–22}, CEM15 may affect Δ vif virions through their RNA components. It is possible that such effects underlie the inability of Δ vif viruses to accumulate reverse transcripts and to establish proviruses in susceptible target cells^{6,23–25}. Notably, the inhibitory action of CEM15 during HIV-1 assembly and virion production stands in contrast to the anti-retroviral effects of *Fv1* and *Ref1*, which act in target cells to restrict infection by incoming murine leukaemia virus (MLV) particles^{26,27}. On the basis of these findings, we propose that CEM15 comprises a significant part (or all) of a form of innate antiviral resistance that acts during the late stages of the viral life cycle, and suggest that relieving its suppression by Vif may lend a fresh perspective to the area of HIV/AIDS therapies. □

Methods

Cells, transfections and gene transfer

All lymphoid and non-lymphoid cell lines were maintained under standard conditions. Peripheral blood mononuclear cells were obtained by venipuncture, stimulated with $5 \mu\text{g ml}^{-1}$ phytohemagglutinin-L and maintained in medium supplemented with interleukin-2. Monolayer cultures of 293T cells were transiently transfected with calcium phosphate, and CEM-SS cells were transduced by a murine retroviral transduction system²⁸ followed by maintenance in medium containing 1 mg ml^{-1} G418.

HIV-1 replication, infectivity and purification

Stocks of wild-type and Δ vif viruses were prepared by transfection of 293T cells, quantified by enzyme-linked immunosorbent assay (ELISA) for p24^{Gag} content and stored at -80°C . Spreading infections were initiated with stocks normalized according to p24^{Gag} levels, the cells were passaged, and replication was monitored over time as the accumulation of p24^{Gag} in the culture supernatants. Single-cycle infectivities were determined by challenging the C8166-CCR5/HIV-CAT indicator cell line (5×10^5 cells) with viruses corresponding to 5 ng p24^{Gag}, and measuring the induced expression of CAT in cell lysates after ~ 28 h (ref. 24). Virions were isolated from culture supernatants by clarification at 500g for 5 min, passage through 0.45- μm filters and pelleting through a 20% (w/v) sucrose cushion at 100,000g for 1.5 h. Pellets were then resuspended in PBS buffer, applied to 20–60% (w/v) sucrose gradients and centrifuged at 200,000g for 2 h. The peak fractions were then identified, and virions were collected by ultracentrifugation at 100,000g for 1.5 h.

Cloning of CEM15 and other plasmids

Preparations of polyadenylated RNA (Qiagen Oligotex mRNA Midi Kit) derived from wild-type CEM or CEM-SS cells infected with HIV-1 ($\sim 50\%$ antigen positive) were reverse transcribed and subjected to subtraction using the PCR-Select system (Clontech), with the CEM sample serving as the 'tester'. Ensuing subcloned cDNA fragments were sequenced and used as probes in northern blotting screens of RNAs isolated from a panel of non-permissive and permissive cells (data not shown). A cDNA encoding the predicted 384-amino-acid open reading frame of CEM15 was derived by PCR with reverse transcription (RT-PCR) and inserted into the *EcoRI* site of pcDNA3.1 (Invitrogen). Three copies of the sequence encoding the influenza virus haemagglutinin (HA) epitope tag were included at the 3' end of CEM15 to create pCEM15:HA.

The wild-type and Δ vif X4 provirus expression vectors, pIIIB and pIIIB/ Δ vif, have been described²⁴, as has the wild-type R5 proviral vector, pYU-2 (ref. 28). A Δ vif derivative of pYU-2 was generated by removal of nucleotides 5,126–5,138. The retroviral vector NG/pU, based on murine stem cell virus (MSCV), was derived from MIGR1 (ref. 29) by replacement of the green fluorescent protein gene (*GFP*) with the neomycin phosphotransferase gene, and the CEM15:HA cassette was then inserted as an *XhoI*–*EcoRI* fragment to create NG/C15. The luciferase expression vector, pLUC, was constructed by insertion of the luciferase gene from pBI-L (Clontech) into pcDNA3.

RNA isolation and northern analyses

Total RNAs were isolated from uninfected or infected cells using standard procedures, and $\sim 5 \mu\text{g}$ was then resolved on MOPS-containing agarose gels and transferred to GeneScreen (NEN Life Science Products). The filters were hybridized to random primed ³²P-labelled DNA probes corresponding to subtracted cDNAs, and visualized by autoradiography.

Immunoblotting and luciferase assays

CEM15:HA, the heterogeneous nuclear (hn)RNP C1/2 proteins, p24^{Gag} and Vif were detected in whole-cell lysates (transfected or transduced cultures) or lysates of purified virions using the 16B12 (Covance), 4F4, 24-3 or 319 monoclonal antibodies³⁰, respectively, for primary hybridization and enhanced chemiluminescence. In some cases, luciferase activity was measured in the same cell lysates using the Luciferase Assay System (Promega).

Received 30 May; accepted 1 July 2002; doi:10.1038/nature00939.

Published online 14 July 2002.

- Desrosiers, R. C. et al. Identification of highly attenuated mutants of simian immunodeficiency virus. *J. Virol.* **72**, 1431–1437 (1998).
- Fisher, A. G. et al. The *src* gene of HIV-1 is required for efficient virus transmission *in vitro*. *Science* **237**, 888–893 (1987).
- Gabuzda, D. H. et al. Role of *vif* in replication of human immunodeficiency virus type 1 in CD4⁺ T lymphocytes. *J. Virol.* **66**, 6489–6495 (1992).



Figure 6 Amino-acid sequence of CEM15. ClustalW alignment showing similarity of the N- and C-terminal domains of CEM15 (AAG14956) (CN and CC respectively), its mouse orthologue (AAH03314) (MN and MC respectively), human phorbolin-1 (AAA03706) (P) and human apobec-1 (NP_001635) (A) proteins; residue numbers for each line of sequence are provided. Asterisks, colons and full stops indicate identical, conserved and semiconserved residues, respectively. The boxed region indicates the Prosite cytosine nucleoside/nucleotide deaminase zinc-binding motif PS00903, with putative zinc-binding histidine and cysteine residues shown in bold.

4. Simon, J. H. M. *et al.* The regulation of primate immunodeficiency virus infectivity by Vif is cell species restricted: a role for Vif in determining virus host range and cross-species transmission. *EMBO J.* **17**, 1259–1267 (1998).
5. Strebel, K. *et al.* The HIV “A” (*sor*) gene product is essential for virus infectivity. *Nature* **328**, 728–730 (1987).
6. von Schwedler, U., Song, J., Aiken, C. & Trono, D. *vif* is crucial for human immunodeficiency virus type 1 proviral DNA synthesis in infected cells. *J. Virol.* **67**, 4945–4955 (1993).
7. Madani, N. & Kabat, D. An endogenous inhibitor of human immunodeficiency virus in human lymphocytes is overcome by the viral Vif protein. *J. Virol.* **72**, 10251–10255 (1998).
8. Simon, J. H. M., Gaddis, N. C., Fouchier, R. A. M. & Malim, M. H. Evidence for a newly discovered cellular anti-HIV-1 phenotype. *Nature Med.* **4**, 1397–1400 (1998).
9. Foley, G. E. *et al.* Loss of neoplastic properties in vitro. II. Observations on KB sublines. *Cancer Res.* **25**, 1254–1261 (1965).
10. Nara, P. L. & Fischinger, P. J. Quantitative infectivity assay for HIV-1 and -2. *Nature* **332**, 469–470 (1988).
11. Hassaine, G. *et al.* The tyrosine kinase Hck is an inhibitor of HIV-1 replication counteracted by the viral vif protein. *J. Biol. Chem.* **276**, 16885–16893 (2001).
12. Camaur, D. & Trono, D. Characterization of human immunodeficiency virus type 1 Vif particle incorporation. *J. Virol.* **70**, 6106–6111 (1996).
13. Fouchier, R. A. M., Simon, J. H. M., Jaffe, A. B. & Malim, M. H. Human immunodeficiency virus type 1 Vif does not influence expression or virion incorporation of *gag*-, *pol*-, and *env*-encoded proteins. *J. Virol.* **70**, 8263–8269 (1996).
14. Liu, H. *et al.* The Vif protein of human and simian immunodeficiency viruses is packaged into virions and associates with viral core structures. *J. Virol.* **69**, 7630–7638 (1995).
15. Teng, B., Burant, C. F. & Davidson, N. O. Molecular cloning of an apolipoprotein B messenger RNA editing protein. *Science* **26**, 1816–1819 (1993).
16. Madsen, P. *et al.* Psoriasis upregulated phorbol-1 shares structural but not functional similarity to the mRNA-editing protein apobec-1. *J. Invest. Dermatol.* **113**, 162–169 (1999).
17. Bhattacharya, S., Navaratnam, N., Morrison, J. R., Scott, J. & Taylor, W. R. Cytosine nucleoside/nucleotide deaminases and apolipoprotein B mRNA editing. *Trends Biochem. Sci.* **19**, 105–106 (1994).
18. MacGinnitie, A. J., Anant, S. & Davidson, N. O. Mutagenesis of apobec-1, the catalytic subunit of the mammalian apolipoprotein B mRNA editing enzyme, reveals distinct domains that mediate cytosine nucleoside deaminase, RNA binding, and RNA editing activity. *J. Biol. Chem.* **270**, 14768–14775 (1995).
19. Smith, A. A., Carlow, D. C., Wolfenden, R. & Short, S. A. Mutations affecting transition-state stabilization by residues coordinating zinc at the active site of cytidine deaminase. *Biochemistry* **33**, 6468–6474 (1994).
20. Dettenhofer, M., Cen, S., Carlson, B. A., Kleiman, L. & Yu, X.-F. Association of human immunodeficiency virus type 1 Vif with RNA and its role in reverse transcription. *J. Virol.* **74**, 8938–8945 (2000).
21. Khan, M. A. *et al.* Human immunodeficiency virus type 1 Vif protein is packaged into the nucleoprotein complex through an interaction with viral genomic RNA. *J. Virol.* **75**, 7252–7265 (2001).
22. Zhang, H., Pomerantz, R. J., Dornadula, G. & Sun, Y. Human immunodeficiency virus type 1 Vif protein is an integral component of an mRNP complex of viral RNA and could be involved in the viral RNA folding and packaging process. *J. Virol.* **74**, 8252–8261 (2000).
23. Courcou, M. *et al.* Peripheral blood mononuclear cells produce normal amounts of defective Vif-human immunodeficiency virus type 1 particle which are restricted for the preretrotranscription steps. *J. Virol.* **69**, 2068–2074 (1995).
24. Simon, J. H. M. & Malim, M. H. The human immunodeficiency virus type 1 Vif protein modulates the postpenetration stability of viral nucleoprotein complexes. *J. Virol.* **70**, 5297–5305 (1996).
25. Sova, P. & Volsky, D. J. Efficiency of viral DNA synthesis during infection of permissive and nonpermissive cells with *vif*-negative human immunodeficiency virus type 1. *J. Virol.* **67**, 6322–6326 (1993).
26. Pryciak, P. M. & Varmus, H. E. *Fv-1* restriction and its effects on murine leukemia virus integration in vivo and in vitro. *J. Virol.* **66**, 5959–5966 (1992).
27. Towers, G. *et al.* A conserved mechanism of retrovirus restriction in mammals. *Proc. Natl Acad. Sci. USA* **97**, 12295–12299 (2000).
28. Fouchier, R. A. M., Meyer, B. E., Simon, J. H. M., Fischer, U. & Malim, M. H. HIV-1 infection of non-dividing cells: evidence that the amino-terminal basic region of the viral matrix protein is important for *Gag* processing but not for post-entry nuclear import. *EMBO J.* **16**, 4531–4539 (1997).
29. Pear, W. S. *et al.* Efficient and rapid induction of a chronic myelogenous leukemia-like myeloproliferative disease in mice receiving P210 bcr/abl-transduced bone marrow. *Blood* **92**, 3780–3792 (1998).
30. Simon, J. H. M. *et al.* The Vif and *Gag* proteins of human immunodeficiency virus type 1 colocalize in infected human T cells. *J. Virol.* **71**, 5259–5267 (1997).

Supplementary Information accompanies the paper on *Nature's* website (<http://www.nature.com/nature>).

Acknowledgements

We thank D. Gabuzda for the CEM cells. This work was supported by Research Grants from the National Institutes of Health (M.H.M. and A.M.S.), the National Science Foundation (N.C.G.) and the UK Medical Research Council (M.H.M.). M.H.M. is an Elizabeth Glaser Scientist supported by the Elizabeth Glaser Pediatric AIDS Foundation.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.H.M. (e-mail: michael.malim@kcl.ac.uk).

Gut hormone PYY₃₋₃₆ physiologically inhibits food intake

Rachel L. Batterham^{*†}, Michael A. Cowley^{†‡§}, Caroline J. Small^{*}, Herbert Herzog^{||}, Mark A. Cohen^{*}, Catherine L. Dakin^{*}, Alison M. Wren^{*}, Audrey E. Brynes^{*}, Malcolm J. Low[§], Mohammad A. Ghatei^{*}, Roger D. Cone[§] & Stephen R. Bloom^{*}

^{*} Imperial College Faculty of Medicine at Hammersmith Campus, Du Cane Road, London W12 0NN, UK

[‡] Oregon National Primate Research Centre, Oregon Health and Sciences University, 505 NW 185th, Beaverton, Oregon 97006, USA

[§] The Vollum Institute, Oregon Health and Sciences University, 3181 SW Sam Jackson Park Road, Portland, Oregon 97201, USA

^{||} Neurobiology Program, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, Sydney, New South Wales 2010, Australia

[†] These authors contributed equally to this work

Food intake is regulated by the hypothalamus, including the melanocortin and neuropeptide Y (NPY) systems in the arcuate nucleus¹. The NPY Y2 receptor (Y2R), a putative inhibitory presynaptic receptor, is highly expressed on NPY neurons² in the arcuate nucleus, which is accessible to peripheral hormones³. Peptide YY₃₋₃₆ (PYY₃₋₃₆), a Y2R agonist⁴, is released from the gastrointestinal tract postprandially in proportion to the calorie content of a meal⁵⁻⁷. Here we show that peripheral injection of PYY₃₋₃₆ in rats inhibits food intake and reduces weight gain. PYY₃₋₃₆ also inhibits food intake in mice but not in Y2R-null mice, which suggests that the anorectic effect requires the Y2R. Peripheral administration of PYY₃₋₃₆ increases c-Fos immunoreactivity in the arcuate nucleus and decreases hypothalamic *Npy* messenger RNA. Intra-arcuate injection of PYY₃₋₃₆ inhibits food intake. PYY₃₋₃₆ also inhibits electrical activity of NPY nerve terminals, thus activating adjacent pro-opiomelanocortin (POMC) neurons⁸. In humans, infusion of normal postprandial concentrations of PYY₃₋₃₆ significantly decreases appetite and reduces food intake by 33% over 24 h. Thus, postprandial elevation of PYY₃₋₃₆ may act through the arcuate nucleus Y2R to inhibit feeding in a gut-hypothalamic pathway.

The orexigenic NPY and the anorectic alpha melanocyte-stimulating hormone (α -MSH) systems of the hypothalamic arcuate nucleus are involved in the central regulation of appetite¹. However, the potential mechanisms that signal meal ingestion directly to these hypothalamic-feeding circuits are unclear. PYY₃₋₃₆ is a gut-derived hormone that is released postprandially in proportion to the calories ingested⁵. We therefore investigated the effects of peripheral administration of PYY₃₋₃₆ on feeding.

An intraperitoneal (i.p.) injection of PYY₃₋₃₆ to freely feeding rats before the onset of the dark phase significantly decreased subsequent food intake (Fig. 1a). A similar inhibition of feeding was seen after i.p. injection in rats fasted for 24 h (Supplementary Information Fig. 1). A time course of the plasma PYY₃₋₃₆ concentrations after i.p. injection of PYY₃₋₃₆ showed a peak at 15 min after injection, which was within the normal postprandial range (peak PYY₃₋₃₆ 15 min after i.p. injection of 0.3 μ g per 100 g (body weight), 99.3 ± 10.4 pmol l⁻¹; peak postprandial PYY₃₋₃₆, 112.1 ± 7.8 pmol l⁻¹; $n = 8-10$ per group), suggesting that physiological concentrations of PYY₃₋₃₆ inhibit feeding. PYY₃₋₃₆ did not affect gastric emptying (percentage of food ingested remaining in the stomach at 3 h (ref. 9): PYY₃₋₃₆, $36 \pm 1.9\%$; saline, $37.4 \pm 1.0\%$; $n = 12$). PYY₃₋₃₆ that was administered i.p. twice daily for 7 d reduced cumulative food intake (7-d cumulative food intake: PYY₃₋₃₆, 187.6 ± 2.7 g; saline, 206.8 ± 2.3 g; $n = 8$ per group, $P < 0.0001$) and decreased body weight gain (PYY₃₋₃₆, 48.2 ± 1.3 g; saline, 58.7 ± 1.9 g; $n = 8$ per group, $P < 0.002$; Fig. 1b).

To investigate whether this inhibition of food intake involved a hypothalamic pathway, we examined c-Fos expression in the arcuate nucleus, which is an important centre of feeding control^{1,8}, after a single i.p. injection of PYY₃₋₃₆. There was a twofold increase in the number of cells positive for c-Fos in the lateral arcuate of the rat (PYY₃₋₃₆, 168 ± 2; saline, 82.7 ± 5; *n* = 3, *P* < 0.0001). Similarly, in *Pomc-EGFP* transgenic mice⁸ i.p. administration of PYY₃₋₃₆ resulted in a 1.8-fold increase in the number of arcuate cells positive for c-Fos (Fig. 1c, d), as compared with saline control animals (PYY₃₋₃₆, 250 ± 40; saline, 137 ± 15; *n* = 5, *P* < 0.05). Injection of i.p. PYY₃₋₃₆ caused a 2.6-fold increase in the proportion of POMC neurons that expressed c-Fos (PYY₃₋₃₆, 20.4 ± 2.9%; saline, 8 ± 1.4%; *n* = 5, *P* < 0.006; Fig. 1e, f).

As these observations suggested that PYY₃₋₃₆ might act through the arcuate nucleus, we studied the effects of PYY₃₋₃₆ on NPY and POMC circuits in the hypothalamus. In view of the sustained

inhibition of food intake and the effects on weight gain after peripheral administration of PYY₃₋₃₆, we measured both *Pomc* and *Npy* hypothalamic messenger RNA (mRNA) using RNase protection assays. A significant decrease in *Npy* mRNA in response to PYY₃₋₃₆ was observed 6 h after i.p. injection, as compared with saline-treated animals (relative optical density units: saline, 17.3 ± 2.0; PYY₃₋₃₆, 8.8 ± 1.0; *P* < 0.02). A nonsignificant increase occurred in *Pomc* mRNA.

PYY₃₋₃₆ shares 70% amino-acid sequence identity with NPY and acts through NPY receptors¹⁰. The Y2R is a putative inhibitory presynaptic receptor and is highly expressed on the arcuate NPY neurons², although not on the neighbouring POMC neurons. PYY₃₋₃₆ is a high affinity agonist at the Y2 receptor⁷. We thought that peripheral PYY₃₋₃₆ might inhibit food intake through the Y2R in the arcuate nucleus, an area that is directly accessible to circulating hormones³.

To investigate this hypothesis, we injected PYY₃₋₃₆ directly into the arcuate nucleus¹¹. In rats fasted for 24 h, food intake was significantly decreased by doses as low as 100 fmol (Fig. 2a), which resulted in a similar inhibition to that seen after i.p. administration. To establish whether these effects occurred through the Y2R we used a Y2R selective agonist¹², Y2A (*N*-acetyl [Leu²⁸, Leu³¹] NPY(24–36)). Its affinity was confirmed using receptor-binding studies¹³ on cell lines that expressed the NPY Y1, Y2 and Y5 receptors (inhibitor concentration for half-maximum response (IC₅₀): Y2, 1.3 ± 0.2 nM; Y1, >5,000 nM; Y5, >5,000 nM). Intra-arcuate nucleus injection of Y2A in rats previously fasted for 24 h dose-dependently inhibited (100 fmol to 1 nmol) food intake (chow ingested 2 h after injection: 0.1 nmol of Y2A, 6.2 ± 0.5 g; saline, 8.2 ± 0.6 g, *n* = 8 per group, *P* < 0.05). To confirm the anatomical specificity of this effect we injected Y2A (100 fmol to 1 nmol) into the paraventricular nucleus (PVN)¹⁴ of rats fasted for 24 h and found no alteration of food intake (2 h after injection: saline, 8.3 ± 0.4 g; 0.1 nmol of Y2A, 8.0 ± 0.6 g; *n* = 8 per group). To define further the role of the Y2R in the feeding inhibition caused by peripheral PYY₃₋₃₆, we examined the effect of PYY₃₋₃₆ on Y2r-null mice and littermate controls. PYY₃₋₃₆ inhibited daytime feeding in a dose-responsive manner in fasted male wild-type mice but did not inhibit food intake in fasted male Y2r-null mice (Fig. 2b, c).

We examined the electrophysiological response of hypothalamic POMC neurons to administration of both PYY₃₋₃₆ and Y2A. These neurons were identified using mice with targeted expression of green fluorescent protein in POMC neurons⁸. PYY₃₋₃₆ disinhibited the POMC neurons, resulting in a significant depolarization of 19 of the 22 POMC neurons tested (Fig. 3a, inset; 10.3 ± 2.1 mV depolar-

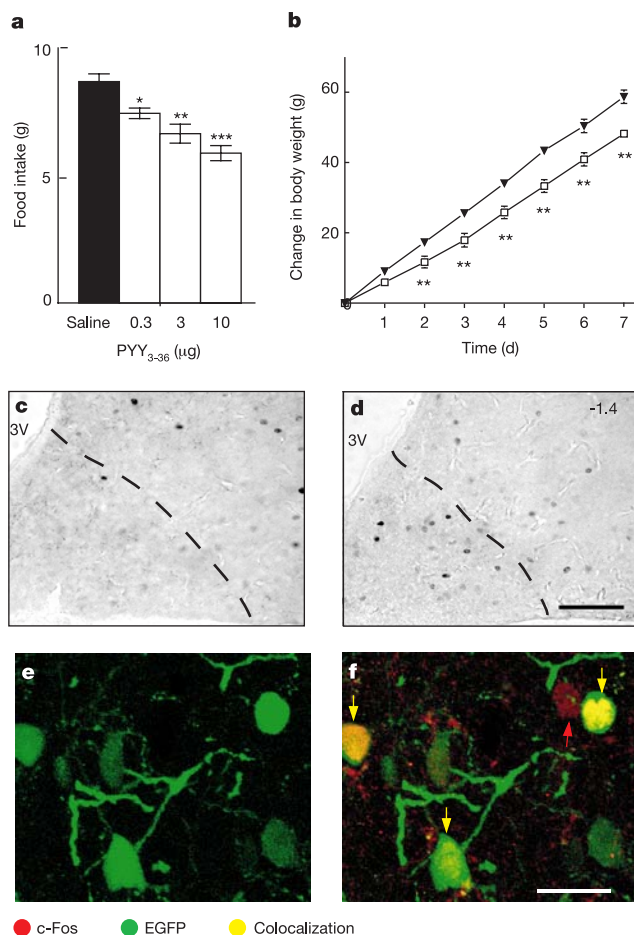


Figure 1 Feeding response to PYY₃₋₃₆ in rats and c-Fos expression in *Pomc-EGFP* mice. **a**, Dark-phase feeding: food intake after i.p. injection of PYY₃₋₃₆. Freely feeding rats were injected with PYY₃₋₃₆ at the doses indicated (μg per 100 g) or saline, just before 'lights off', and 4-h cumulative food intake was measured. Results are means ± s.e.m. (*n* = 8 per group); asterisk, *P* < 0.05; double asterisk, *P* < 0.01; triple asterisk, *P* < 0.001 versus saline. **b**, Body weight gain during chronic treatment with PYY₃₋₃₆. Rats were injected i.p. with PYY₃₋₃₆ (5 μg per 100 g; open squares) or saline (filled inverted triangles) twice daily for 7 d. Body weight gain was calculated each day. Results are expressed as means ± s.e.m. (*n* = 12 per group); double asterisk, *P* < 0.01 versus saline. **c**, **d**, Representative example (Bregma, 1.4 mm)²² of c-Fos expression in the arcuate nucleus of *Pomc-EGFP* mice in response to i.p. injection of saline (**c**) or PYY₃₋₃₆ (5 μg per 100 g; **d**). Scale bar, 100 μm. 3V, third ventricle. **e**, **f**, Example of POMC-EGFP neurons (**e**), and c-Fos immunoreactivity (**f**) either colocalizing (yellow arrows) or alone (red arrow). Scale bar, 25 μm.

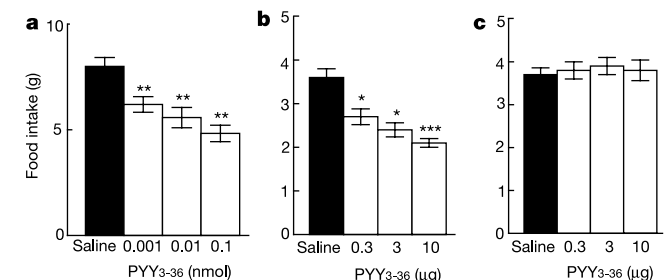


Figure 2 Feeding effects of intra-arcuate injection of PYY₃₋₃₆ in rats and of i.p. PYY₃₋₃₆ in Y2r-null mice. **a**, Food intake after intra-arcuate injection of PYY₃₋₃₆. Fasted rats were injected with saline or PYY₃₋₃₆ into the arcuate nucleus at the doses indicated, and 2-h cumulative food intake was measured; double asterisk, *P* < 0.01 versus saline. **b**, **c**, Feeding response to PYY₃₋₃₆ in Y2r-null mice after i.p. administration. Wild-type littermates (**b**) and Y2r-null mice (**c**), fasted for 24 h, were injected with PYY₃₋₃₆ at the doses indicated (μg per 100 g) or saline, and 4-h cumulative food intake was measured. Results are the mean ± s.e.m. (*n* = 5 per group); asterisk, *P* < 0.05; double asterisk, *P* < 0.01; triple asterisk, *P* < 0.001 versus saline.

ization, $n = 22$, $P < 0.0003$). A similar depolarization was seen with Y2A (8.7 ± 1.8 mV depolarization, $n = 9$, $P < 0.002$). The depolarization caused by PYY₃₋₃₆ stimulated a significant increase in the frequency of action potentials in POMC neurons (Fig. 3a; 93% increase over control, $P < 0.05$, $n = 22$). In the whole-cell mode, the effect of PYY₃₋₃₆ was sometimes reversed on wash-out, but only after a long latency (30 min). We have observed a similar effect on wash-out of leptin in these neurons.

To exclude effects of cellular rundown or seal deterioration, we examined the effects of PYY₃₋₃₆ in the 'loose cell-attached' (or extracellular) configuration. PYY₃₋₃₆ caused a reversible fivefold increase in the frequency of action potentials in loose cell-attached recordings of POMC neurons (Fig. 3b). This increase in firing rate occurred with the same latency, because PYY₃₋₃₆ reduced the frequency of inhibitory postsynaptic currents (IPSCs) onto all 13 POMC neurons tested (Fig. 3c; $51.9 \pm 9.2\%$ reduction, $n = 13$, $P < 0.0001$), which indicated a reduced frequency of GABA (γ -aminobutyric acid) release onto POMC neurons. Notably, the firing rate of POMC neurons returned to basal levels, in spite of continued inhibition of IPSCs. A similar effect on IPSC frequency was seen with Y2A ($44.4 \pm 9.3\%$ reduction, $n = 8$, $P < 0.004$), which suggests that this effect occurs through Y2R. PYY₃₋₃₆ (25 nM) caused a hyperpolarization (5.2 ± 1.16 mV; $P < 0.004$, $n = 5$) of unidentified, but presumably NPY-containing, non-POMC neurons in the arcuate nucleus (data not shown). There is a tonic GABA-mediated inhibition of POMC neurons by NPY neurons⁸, and these results suggest that PYY₃₋₃₆ acts by inhibiting NPY neurons, thus decreasing this GABA-mediated tone and consequentially disinhibi-

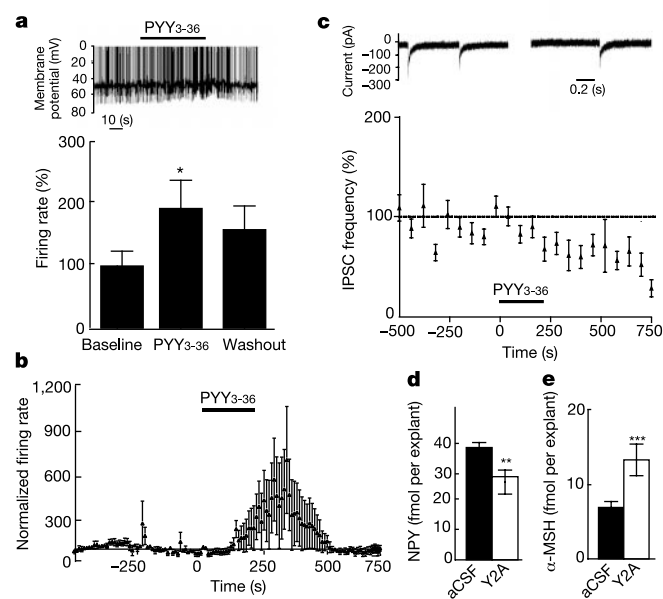


Figure 3 Electrophysiological and neuropeptide responses to PYY₃₋₃₆ and Y2A. **a**, Effect of PYY₃₋₃₆ (10 nM) on the frequency of action potentials in POMC neurons (whole-cell configuration recordings; $n = 22$); asterisk, $P < 0.05$. PYY₃₋₃₆ was administered at time 0 for 3 min; baseline, -3 to 0 min; PYY₃₋₃₆, 2 – 5 min; and wash-out, 8 – 11 min. The top panel shows a representative recording of membrane potential and action potential frequency. **b**, Effect of PYY₃₋₃₆ (10 nM) on the frequency of action potentials in loose cell-attached patch recordings ($n = 8$). Data from individual cells were normalized to the firing rate for the 200 s before PYY₃₋₃₆ addition. **c**, Effect of PYY₃₋₃₆ (50 nM) on spontaneous IPSCs onto POMC neurons ($n = 13$). The top panel shows a representative recording of IPSCs before and after PYY₃₋₃₆ (50 nM), respectively. Results in **a**–**c** are expressed as means $\pm$ s.e.m. **d**, **e**, NPY (**d**) and α -MSH (**e**) released from hypothalamic explants in response to Y2A. Hypothalamic slices were incubated with artificial CSF (aCSF), with or without 50 nM Y2A, for 45 min. Results are expressed as means $\pm$ s.e.m. ($n = 40$); double asterisk, $P < 0.01$; triple asterisk, $P < 0.001$ versus saline.

biting POMC neurons. The effect of Y2A on peptide secretion was also examined using hypothalamic explants¹⁴. Y2A significantly decreased NPY release, with a concomitant increase in α -MSH release from hypothalamic explants (Fig. 3d, e). Together, these observations suggest that PYY₃₋₃₆ modulates both the NPY and melanocortin systems in the arcuate nucleus.

Because of the known importance of the melanocortin system in man¹⁵, and the profound effects of PYY₃₋₃₆ on both feeding and weight change observed in rodents, we investigated the effects of PYY₃₋₃₆ on appetite and food intake in man. Twelve healthy fasted, non-obese volunteers (six men and six women: mean age, 26.7 ± 0.7 years; BMI, 24.6 ± 0.94 kg m⁻²) were infused with PYY₃₋₃₆ (0.8 pmol per kg (body weight) per min) or saline for 90 min in a double-blind placebo-controlled crossover study. PYY₃₋₃₆ plasma increased from a mean basal concentration of 8.3 ± 1.0 pM to 43.5 ± 3 pM during the PYY₃₋₃₆ infusion and mimicked postprandial concentrations^{5,6}. After infusion, PYY₃₋₃₆ concentrations returned to basal levels within 30 min. PYY₃₋₃₆ infusion resulted in a significant decrease in hunger scores¹⁶ (Fig. 4c). Calorie intake during a free-choice buffet meal¹⁷ 2 h after the termination of the infusion was reduced by more than a third as compared with saline controls ($36 \pm 7.4\%$, $P < 0.0001$; Fig. 4a). There was no effect on fluid intake and no difference in sensations of fullness or nausea reported by the volunteers (data not shown). PYY₃₋₃₆ administration had no effect on gastric emptying, as estimated by the paracetamol absorption method^{17,18}, or on plasma glucose, plasma leptin or insulin (data not shown). Analysis of the food diaries showed a significant inhibition of food intake in the 12-h period after the PYY₃₋₃₆ infusion (saline, $2,205 \pm 243$ kcal; PYY₃₋₃₆, $1,474 \pm 207$ kcal). But food intake between the two groups during the 12-h to 24-h period was virtually identical. Overall

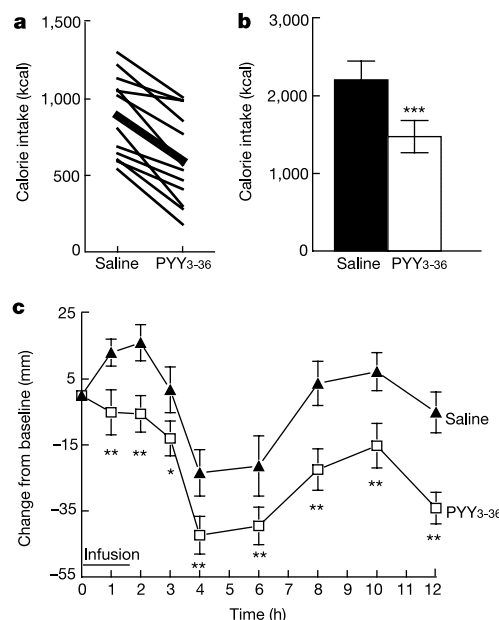


Figure 4 Effect of PYY₃₋₃₆ infusion on appetite and food intake in humans. **a**, Calorie intake from free-choice buffet meal 2 h after infusion with saline or PYY₃₋₃₆. The thin lines indicate individual changes in calorie intake for each subject between saline and PYY₃₋₃₆ administration. The thick line represents mean change between the two infusions ($n = 12$). **b**, Total calorie intake, as assessed by food diaries, for the 24-h period after either saline or PYY₃₋₃₆ infusion. Data are expressed means $\pm$ s.e.m. ($n = 12$); triple asterisk, $P < 0.0001$ versus saline. **c**, Appetite score (relative scale). Visual analogue scores¹⁶ show perceived hunger during and after infusions. The results are presented as change from baseline scores and are the mean $\pm$ s.e.m. for all 12 subjects. Asterisk, $P < 0.05$; double asterisk, $P < 0.01$ versus saline.

there was a 33% decrease in cumulative total calorie consumption in the 24-h period after the PYY₃₋₃₆ infusion (Fig. 4b). These findings show that infusing amounts of PYY₃₋₃₆ that match postprandial concentrations cause a marked inhibition of both appetite and food intake in man.

It has been proposed that the cells in the arcuate nucleus detect circulating peripheral satiety signals and relay these signals to other brain regions¹⁹. This is supported by the observation that leptin modifies the activity of both the POMC and NPY arcuate neurons⁸. We have now shown, through a combination of electrophysiological and hypothalamic explant studies, that the gut hormone, PYY₃₋₃₆, can directly influence hypothalamic circuits, which results in coordinate changes in POMC and NPY action. In addition, PYY₃₋₃₆ administered directly into this brain region reduces food intake. Our data show that postprandial concentrations of PYY₃₋₃₆ inhibit food intake in both rodents and man for up to 12 h, which suggests that PYY₃₋₃₆ has a role in 'longer term' regulation of food intake. This contrasts with previously characterized gut-derived 'short-term' satiety signals such as cholecystokinin^{1,20}, the effects of which are relatively short-lived. The failure of PYY₃₋₃₆ to inhibit food intake in the Y2r-null mice provides further evidence that PYY₃₋₃₆ reduces food intake through a Y2R-dependent mechanism. Our results suggest that a gut-hypothalamic pathway that involves postprandial PYY₃₋₃₆ acting at the arcuate Y2R has a role in regulating feeding. Thus, the PYY₃₋₃₆ system may provide a therapeutic target for the treatment of obesity. □

Methods

Animals

We maintained male Wistar rats (200–250 g), aged 7–8 weeks (Charles River Laboratories), under controlled temperature (21–23 °C) and light conditions (lights on 7:00–19:00) with *ad libitum* access to water and food (RM1 diet; SDS) except where stated. We carried out arcuate and paraventricular nuclei cannulations and injections as described^{11,13,14}. Correct intranuclear cannula placement was confirmed histologically at the end of each study period^{11,13,14}. All animal procedures were approved under the British Home Office Animals (Scientific Procedures) Act, 1986. All injection studies on fasting animals were done in the early light phase (8:00 to 9:00). All dark-phase feeding studies injections were done just before lights off.

Male *Pomc-EGFP* mice⁸ were studied at 5–6 weeks of age. Y2r-null mice were generated using Cre-loxP mediated recombination, which results in germline deletion of the whole coding region of the Y2 receptor. We maintained all Y2r-null mice on a mixed C57BL/6-129Sv background. Male mice aged 8–12 weeks with 20–30 g body weight were kept under controlled temperature (21–23 °C) and light conditions (lights on 6:00–18:00) with *ad libitum* access to water and food (Gordon's Speciality Stock feeds) except where stated. All studies were carried out in the early light phase (7:00–8:00).

Intraperitoneal injections

Rats were accustomed to i.p. injection by injections of 0.5 ml of saline for 2 d before the study. For all studies, animals received an i.p. injection of either PYY₃₋₃₆ or saline in 500 µl (for rats) or 100 µl (for mice).

Electrophysiology

Whole-cell patch-clamp recordings were made from POMC neurons in the hypothalamus of 180-µm thick coronal slices from *Pomc-EGFP* mice, as described⁸. Loose cell-attached recordings were made by using extracellular buffer in the electrode solution and by maintaining seal resistance between 3 and 5 MΩ throughout the recording. Firing rates were analysed using mini-analysis protocols (MiniAnalysis, Jaen Software). Vehicle controls were used in this system that had been validated for the electrophysiological actions of neuropeptides⁸. We analysed data by analysis of variance (ANOVA), Neuman-Keuls posthoc comparison and Wilcoxon signed rank test.

Hypothalamic explants

Male Wistar rats were killed by decapitation. The whole brain was removed immediately, mounted with the ventral surface uppermost and placed in a vibrating microtome (Biorad, Microfield Scientific Ltd). A 1.7-mm slice was taken from the base of the brain to include the PVN and the ARC, and immediately transferred to 1 ml of artificial CSF (aCSF)¹⁴ equilibrated with 95% O₂/5% CO₂ and maintained at 37 °C. After an initial 2-h equilibration period, with aCSF replaced every 60 min, the hypothalamus were incubated for 45 min in 600 µl of aCSF (basal period) before being exposed to the Y2A (50 nM) in 600 µl of aCSF. We verified the viability of the tissue by a 45-min exposure to 56 mM KCl; isotonicity was maintained by substituting K⁺ for Na⁺. At the end of each period, the aCSF was removed and frozen at –20 °C until being assayed for NPY and α-MSH by radioimmunoassay.

c-Fos expression

c-Fos expression was measured in adult Wistar rats and *Pomc-EGFP* mice 2 h after i.p. administration of saline or PYY₃₋₃₆ (5 µg per 100 g (body weight)) using standard immunohistochemical techniques²¹. We obtained data from three rats and five mice in each group. For the *Pomc-EGFP* mice, five anatomically matched arcuate nucleus sections²² were counted from each animal, and images were acquired using a Leica TSC confocal microscope²³.

RNase protection assay

We extracted total RNA from hypothalamus in Trizol (Gibco-BRL). Assays were done with an RPAIII kit (Ambion) using 5 µg of RNA and probes specific for NPY, α-MSH and β-actin (internal standard). For each neuropeptide, the ratio of the optical density of the band of neuropeptide mRNA to that of β-actin was calculated. Neuropeptide mRNA is expressed relative to that of saline controls (means ± s.e.m, *n* = 4 per group). For statistical analysis, we used ANOVA with Bonferroni posthoc analysis.

Plasma assays

Human leptin was measured using a commercially available radioimmunoassay (Linco Research). All other plasma hormones were measured using established in-house radioimmunoassays¹⁸. Glucose concentrations were measured using a YSI 2300STAT analyser (Yellow Springs Instruments). We measured concentrations of plasma paracetamol by an enzymatic colorimetric assay (Olympus AU600 analyser).

Human studies

We purchased PYY₃₋₃₆ from Bachem. The Limulus amoebocyte lysate assay test for pyrogen was negative and the peptide was sterile on culture. Ethical approval was obtained from the Local Research Ethics Committee (project registration 2001/6094) and the study was carried out in accordance with the principles of the Declaration of Helsinki. Subjects gave informed written consent. Each subject was studied on two occasions with at least 1 week between each study. Volunteers filled out a food diary for 3 d before each infusion and for the following 24 h. All subjects fasted and drank only water from 20:00 on the evening before each study. Subjects arrived at 8:30 on each study day, were cannulated and then allowed to relax for 30 min before the onset of the study protocol. We collected blood samples every 30 min into heparin-coated tubes containing 5,000 Kallikrein inhibitor units (0.2 ml) of aprotinin (Bayer). Plasma was separated by centrifugation and stored at –70 °C until analysis. Subjects were infused with either saline or 0.8 pmol per kg (body weight) per min PYY₃₋₃₆ for 90 min, in a double-blind randomized crossover design. Two hours after terminating the infusion, subjects were offered an excess free-choice buffet meal¹⁷, such that all appetites could be satisfied. Food and water were weighed pre- and postprandially, and caloric intake was calculated. Appetite ratings were made on 100-mm visual analogue scores with the text expressing the most positive and the negative rating anchored at each end¹⁶. Visual analogue score was used to assess hunger, satiety, fullness, prospective food consumption and nausea. Caloric intake after saline and PYY₃₋₃₆ was compared using a paired *t*-test. The postprandial response curves were compared by ANOVA using repeated paired measures, with time and treatment as factors.

Received 9 January; accepted 13 May 2002; doi:10.1038/nature00887.

- Schwartz, M. W., Woods, S. C., Porte, D. Jr, Seeley, R. J. & Baskin, D. G. Central nervous system control of food intake. *Nature* **404**, 661–671 (2000).
- Broberger, C., Landry, M., Wong, H., Walsh, J. N. & Hökfelt, T. Subtypes of Y1 and Y2 of the neuropeptide Y receptor are respectively expressed in pro-opiomelanocortin and neuropeptide-Y-containing neurons of the rat hypothalamic arcuate nucleus. *Neuroendocrinology* **66**, 393–408 (1997).
- Kalra, S. P. et al. Interacting appetite-regulating pathways in the hypothalamic regulation of body weight. *Endocr. Rev.* **20**, 68–100 (1999).
- Keire, D. A. et al. Primary structures of PYY, [Pro³⁴] PYY and PYY–(3–36) confer different conformations and receptor selectivity. *Am. J. Physiol. Gastrointest. Liver Physiol.* **279**, G126–G131 (2000).
- Pedersen-Bjergaard, U. et al. Influence of meal composition on postprandial peripheral plasma concentrations of vasoactive peptides in man. *Scand. J. Clin. Lab. Invest.* **56**, 497–503 (1996).
- Adrian, T. E. et al. Human distribution and release of a putative new gut hormone, peptide YY. *Gastroenterology* **89**, 1070–1077 (1985).
- Grandt, D. et al. Two molecular forms of peptide YY (PYY) are abundant in human blood: characterisation of a radioimmunoassay recognising PYY₁₋₃₆ and PYY₃₋₃₆. *Regul. Pept.* **51**, 151–159 (1994).
- Cowley, M. A. et al. Leptin activates the anorexigenic POMC neurons through a neural network in the arcuate nucleus. *Nature* **411**, 480–484 (2001).
- Barrachina, M. D., Martinez, V., Wei, J. Y. & Taché, Y. Leptin-induced decrease in food intake is not associated with changes in gastric emptying in lean mice. *Am. J. Physiol.* **272**, R1007–R1011 (1997).
- Soderberg, C. et al. Zebrafish genes for neuropeptide Y and peptide YY reveal origin by chromosome duplication from an ancestral gene linked to the homeobox cluster. *J. Neurochem.* **75**, 908–918 (2000).
- Kim, M. S. et al. Hypothalamic localization of the feeding effect of agouti-related peptide and alpha-melanocyte-stimulating hormone. *Diabetes* **49**, 177–182 (2000).
- Potter, E. K. et al. A novel neuropeptide Y analog N-acetyl [Leu²⁸, Leu³¹] neuropeptide Y–(24–36), with functional specificity of the presynaptic (Y₂) receptor. *Eur. J. Pharmacol.* **267**, 253–262 (1994).
- Small, C. J. et al. Peptide analogue studies of the hypothalamic neuropeptide Y receptor mediating pituitary adrenocorticotrophic hormone release. *Proc. Natl Acad. Sci. USA* **94**, 11686–11691 (1997).
- Kim, M. S. et al. The central melanocortin system affects the hypothalamo-pituitary thyroid axis and may mediate the effect of leptin. *J. Clin. Invest.* **105**, 1005–1011 (2000).
- Barsh, G. S., Farooqi, I. S. & O'Rahilly, S. Genetics of body-weight regulation. *Nature* **404**, 644–651 (2000).
- Raben, A., Tagliabue, A. & Astrup, A. The reproducibility of subjective appetite scores. *Br. J. Nutr.* **73**, 517–530 (1995).
- Edwards, C. M. et al. Exendin-4 reduces fasting and postprandial glucose and decreases energy intake in healthy volunteers. *Am. J. Physiol. Endocrinol. Metab.* **281**, E155–E166 (2001).

18. Tarling, M. M. *et al.* A model of gastric emptying using paracetamol absorption in intensive care patients. *Intens. Care Med.* **23**, 256–260 (1997).
19. Butler, A. A. *et al.* Melanocortin-4 receptor is required for acute homeostatic responses to increased dietary fat. *Nature Neurosci.* **4**, 605–611 (2001).
20. Moran, T. H. Cholecystokinin and satiety: current perspectives. *Nutrition* **16**, 858–865 (2000).
21. Hoffman, G. E., Smith, M. S. & Verbalis, J. G. cFos and related immediate early gene products as markers of activity in neuroendocrine systems. *Front. Neuroendocrinol.* **14**, 173–213 (1993).
22. Franklin, K. B. J. & Paxinos, G. *The Mouse Brain in Stereotaxic Coordinates* (Academic, San Diego, 1997).
23. Grove, K. L., Campbell, R. E., Ffrench-Mullen, J. M., Cowley, M. A. & Smith, M. S. Neuropeptide Y Y5 receptor protein in the cortical/limbic system and brainstem of the rat: expression on gamma-aminobutyric acid and corticotropin-releasing hormone neurons. *Neuroscience* **100**, 731–740 (2000).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements

We thank D. Withers, J. Gardiner, J. Smart, D. Morgan, A. Sainsbury, J. Brundage, J. Williams, K. Takahashi, K. Grove, A. Kennedy, H. Cox and the Hammersmith hypothalamic team for assistance; and G. Williams for providing the antibody to NPY.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.R.B. (e-mail: s.bloom@ic.ac.uk).

A naturally occurring MTA1 variant sequesters oestrogen receptor- α in the cytoplasm

Rakesh Kumar*, Rui-An Wang*, Abhijit Mazumdar*, Amjad H. Talukder*, Mahitosh Mandal*, Zhibo Yang*, Rozita Bagheri-Yarmand*, Aysegul Sahin†, Gabriel Hortobagyi‡, Liana Adam*, Christopher J. Barnes* & Ratna K. Vadlamudi*

* Departments of Molecular and Cellular Oncology, † Pathology, and ‡ Breast Medical Oncology, The University of Texas M. D. Anderson Cancer Center, Houston, Texas 77030, USA

Oestrogen receptor (ER) is a good prognostic marker for the treatment of breast cancers. Upregulation of metastatic tumour antigen 1 (MTA1) is associated with the invasiveness and metastatic potential of several human cancers^{1,2} and acts as a co-repressor of nuclear ER- α ³. Here we identify a naturally occurring short form of MTA1 (MTA1s) that contains a previously unknown sequence of 33 amino acids with an ER-binding motif, Leu-Arg-Ile-Leu-Leu (LRILL). MTA1s localizes in the cytoplasm, sequesters ER in the cytoplasm, and enhances non-genomic responses of ER. Deleting the LRILL motif in MTA1s abolishes its co-repressor function and its interaction with ER, and restores nuclear localization of ER. Dysregulation of human epidermal growth factor receptor-2 in breast cancer cells enhances the expression of MTA1s and the cytoplasmic sequestration of ER. Expression of MTA1s in breast cancer cells prevents ligand-induced nuclear translocation of ER and stimulates malignant phenotypes. MTA1s expression is increased in human breast tumours with no or low nuclear ER. The regulation of the cellular localization of ER by MTA1s represents a mechanism for redirecting nuclear receptor signalling by nuclear exclusion.

By screening a human mammary gland complementary DNA library with a MTA1 cDNA probe, we isolated MTA1 cDNAs of

varying lengths. We sequenced and subcloned the MTA1 cDNAs into the pcDNA3.1 expression vector using the MTA1 cDNA open reading frames (Fig. 1a). To assess the functionality of these clones, we translated MTA1 cDNAs *in vitro* and resolved the resulting protein products on an SDS polyacrylamide gel (Fig. 1b). All of the MTA1 clones except the MTA1-S2 clone were translated into proteins of the expected sizes. The MTA1-S2 clone, which lacks the amino-terminal 58 amino acids and an internal sequence of 47 base pairs (bp), was translated into a protein with a much lower relative molecular mass (M_r) of 44,000 (44K) than that expected on the basis of the MTA1-S2 cDNA length (M_r 78K). We called this much shorter MTA1 variant MTA1s. By re-screening the human mammary gland cDNA library and using recombination methods, we isolated a full-length MTA1s. The T7-tagged MTA1s was translated *in vivo* into a protein of M_r 54K (Fig. 1c).

To determine whether the full-length MTA1s cDNA existed

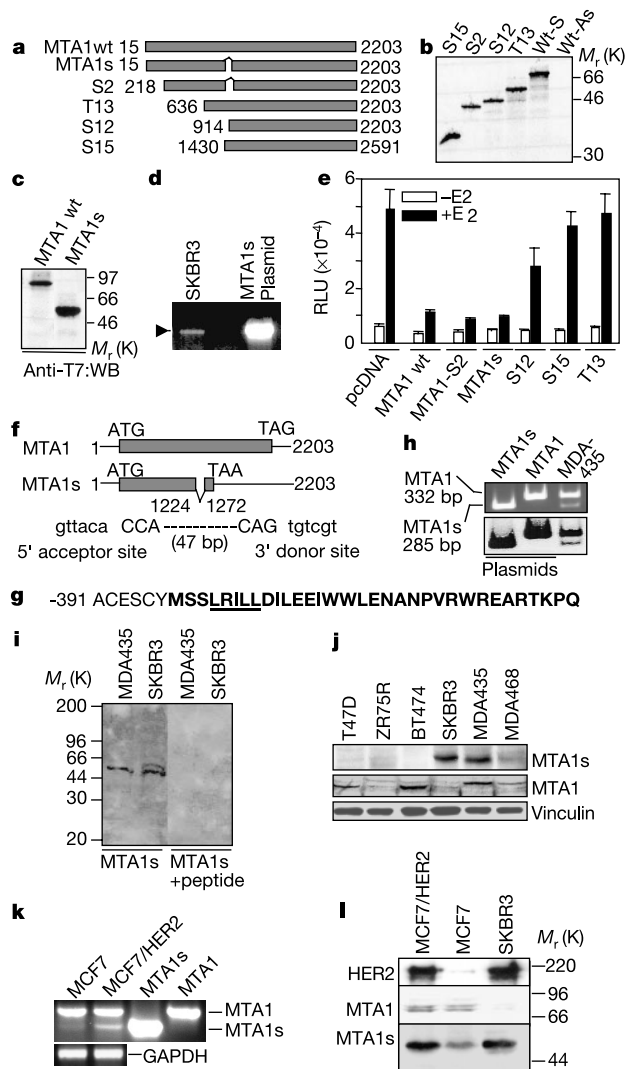


Figure 1 MTA1s is a variant of MTA1. **a**, Representation of the MTA1 clones. **b**, Translation of T7-tagged MTA1 clones. **c**, Expression of T7-MTA1 and T7-MTA1s cDNA. **d**, RT-PCR amplification of the MTA1s open reading frame. **e**, Repression of ERE-luc activity by MTA1 clones. **f**, Human MTA1s transcript generated by alternative splicing. **g**, MTA1s sequence showing the unique 33 amino acids (bold). **h**, RT-PCR (top) and Southern blot (bottom) analysis of RNA from MDA-MB435 cells. **i**, Characterization of MTA1s antibody. **j**, **k**, Expression of MTA1s protein and MTA1s mRNA by immunoblotting and RT-PCR. **l**, Expression of HER2, MTA1 and MTA1s in indicated cell lines.

naturally, we created a MTA1s-specific cDNA library that was based on the reverse transcription of RNA isolated from SKBR3 breast cancer cells. Polymerase chain reaction (PCR) analysis with MTA1s-specific primers yielded an expected 1.3-kilobase band of MTA1s DNA (Fig. 1d). Cloning and sequencing of the 1.3-kb band identified several identical clones with a complete open reading frame and identical sequence to that of MTA1s (data not shown), which showed the endogenous presence of MTA1s.

Because MTA1 is a co-repressor of ER³, we investigated the ability of MTA1 cDNAs to repress the transactivating function of ER. The transient expression of MTA1-S2, MTA1s and full-length MTA1 (but not other MTA1 cDNAs) in MCF-7 cells effectively blocked oestradiol stimulation of oestrogen responsive element (ERE) transcription (Fig. 1e). MTA1s also inhibited transcription regulated by the glucocorticoid responsive element and the progesterone responsive element, but not the retinoic-acid responsive element (see Supplementary Information).

Analysis of the MTA1-S2 sequence indicated a deletion of 47 bases between nucleotides 1,224 and 1,272 of MTA1, which was generated by alternative splicing (Fig. 1f). Sequence evaluation showed that the 3' donor site has a consensus site for splicing^{4,5}, whereas the 5' site has a non-consensus splicing acceptor site. This deletion causes a frameshift that leads to the addition of a unique sequence of 33 amino acids containing a potential nuclear receptor-binding motif Leu-Arg-Ile-Leu-Leu (LRILL; Fig. 1g). A murine homologue of MTA1s also exists, and we isolated a mouse genomic fragment containing the splicing region by screening a mouse bacterial artificial chromosome library using a mouse probe (from the National Institutes of Health) corresponding to nucleo-

tides 1,782 to 2,051 of the MTA1 cDNA. The structure of intron and exon boundaries suggested that a cryptic splice site in exon 14 is responsible for generating MTA1s (see Supplementary Information).

To resolve the 47-bp difference between the MTA1 and MTA1s transcripts, we carried out Southern blot analysis after PCR with reverse transcription (RT-PCR) of the 1,023–1,355 region using RNA extracted from MDA-MB435 breast cancer cells. As expected, there were two distinct bands corresponding to products generated from the MTA1 and MTA1s plasmids (Fig. 1h). The sequence of the lower DNA band was identical to MTA1s. We also confirmed the expression of MTA1s in cancer cells by RNase protection assay (Supplementary Information).

We raised a polyclonal antibody against amino acids 408–429 of the MTA1s sequence and showed that this antibody specifically recognized an endogenous *M_r* 54K MTA1s protein band. This protein was present in variable amounts in breast cancer cell lines and showed different expression to that of MTA1 (Fig. 1i, j). Notably, the deregulation of *HER2* oncogene in a well-characterized

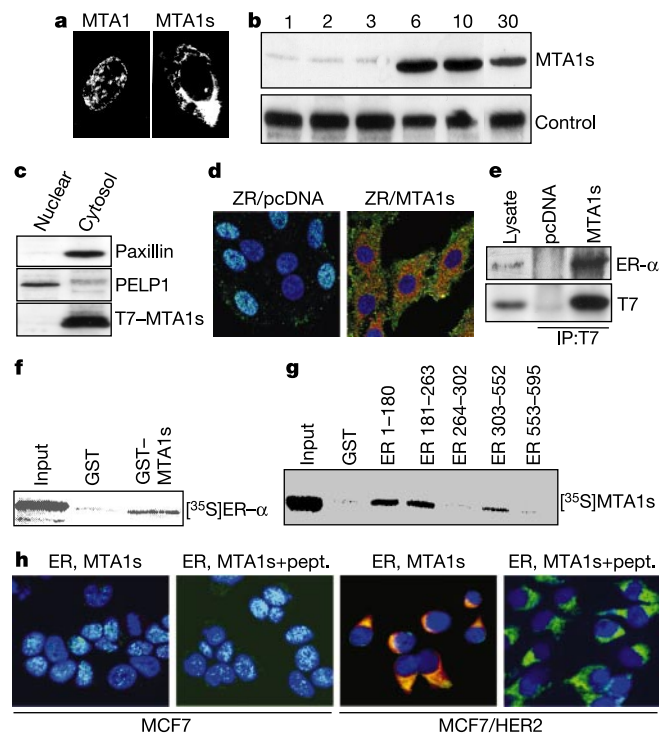


Figure 2 MTA1s is cytoplasmic and interacts with ER. **a**, Confocal microscopy of MCF-7 cells transfected with T7-MTA1 or T7-MTA1s. **b**, Expression of T7-MTA1s in ZR75R clones. **c**, Protein expression in nuclear and cytoplasmic fractions. **d**, Confocal microscopy showing expression of T7-MTA1s (red) and ER-α (green) in ZR75R clones treated with E2 (10⁻⁹ M, 30 min). DNA, blue. **e**, Lysates from ZR75R clones were immunoprecipitated and immunoblotted as indicated. **f**, GST pull-down assay. **g**, MTA1s interactions with GST fusion proteins of the ER domains. **h**, Expression of endogenous MTA1s (red) and ER-α (green) observed by confocal microscopy. DNA, blue.

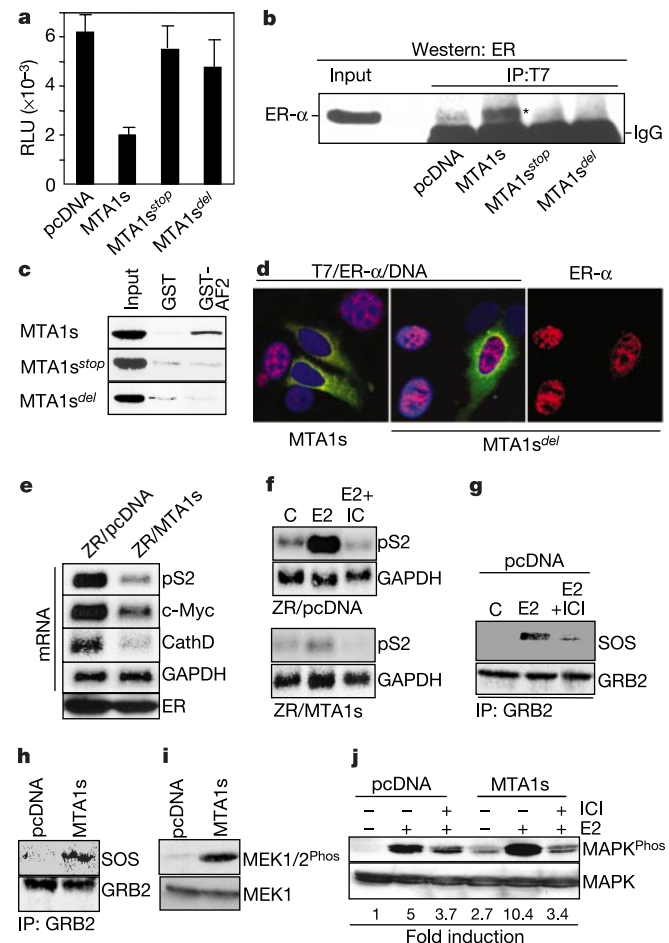


Figure 3 MTA1s LXXLL motif required for interaction and sequestration of ER. **a**, Effect of MTA1s deletions on E2-dependent ERE-luc activity. **b**, Immunoprecipitation of MCF-7 cells lysates expressing T7-MTA1s deletion constructs and immunoblotted for ER-α. **c**, GST pull-down assay. **d**, MCF-7 cells transfected with T7-MTA1s or MTA1s^{del} were treated with E2 and/or ICI 182780 (10⁻⁹ M, 30 min). T7, green; ER, red; DNA, blue. **e**, Expression of E2-regulated genes. **f**, Effect of treatment with E2 (16 h) on pS2 mRNA levels. **g**, **h**, Effect of E2 (10⁻⁹ M, 5 min) with or without ICI 182780 (10⁻⁸ M) on GRB2-SOS. Lysates were immunoprecipitated with an antibody against GRB2 and immunoblotted as shown. **i**, **j**, Effect of E2 (10⁻⁹ M, 5 min) with or without ICI 182780 (10⁻⁸ M) on the p42/44 MAPK pathway. Cell lysates were immunoblotted as shown.

clone of ER-positive MCF-7 cells with aggressive phenotypes^{6,7} was accompanied by an upregulation of MTA1s messenger RNA (Fig. 1k) and protein (Fig. 1l). Specific expression of MTA1s was also determined by confocal scanning microscopy in SKBR-3 cells (see Supplementary Information). MTA1s protein expression and electrophoretic motility were differentially regulated as a function of developmental stage in the mammary gland and were quantitatively distinct from those of MTA1 in murine organs (see Supplementary Information). This differential regulation suggests that the function of MTA1s may be different from that of MTA1 and also that MTA1s is a naturally occurring protein.

As MTA1s lacks a nuclear localization sequence, we examined its subcellular localization and function. Confocal microscopy of transfected cells using a monoclonal antibody against T7 showed that T7-tagged MTA1s localized in the cytoplasm, whereas T7-MTA1 localized in the nucleus as expected (Fig. 2a). To examine the role of MTA1s in breast cancer cells, we next generated stable clones of ZR-75R cells that expressed T7-MTA1s protein (Fig. 2b) and used clone 10 (ZR/MTA1s) in subsequent studies. Subcellular fractionation of the ZR/MTA1s confirmed that T7-MTA1s was localized in the cytoplasm. The purity of the fractions was confirmed by immunoblotting for the nuclear coactivator PELP1 (ref. 8) and cytosolic paxillin (Fig. 2c). We next examined the effect of oestradiol (E2) on the interaction of MTA1s with ER in ZR-75R cells expressing either MTA1s or the empty pcDNA vector. E2 rapidly redistributed ER to the nucleus in ZR75/pcDNA cells but not in more than 90% of the oestradiol-activated ZR75R/MTA1s cells, in which cytoplasmic MTA1s was colocalized with endogenous ER (Fig. 2d).

We confirmed the physical interaction of MTA1s with endogenous ER by immunoprecipitation of T7-MTA1s with a monoclonal antibody against T7 (Fig. 2e). We determined that the association between MTA1s and ER was direct and not mediated by other proteins by showing that *in vitro*-translated ER⁹ protein binds with glutathione S-transferase (GST)-MTA1s but not with GST alone in GST pull-down assays (Fig. 2f). In a reverse experiment, MTA1s interacted with the AF1 (ER amino acids 1–180), DNA-binding (amino acids 181–263) and AF2 (amino acids 303–552) domains of ER (ref. 10 and Fig. 2g). We also used a yeast-based two-hybrid assay to show that full-length MTA1s is required for the functional

interaction of MTA1s and ER *in vivo* (see Supplementary Information). The observed increased expression of the endogenous MTA1s in MCF7/HER2 cells was associated with a significant redistribution of endogenous ER to the cytoplasm (Fig. 2h). In addition to ER, MTA1s also sequestered other members of type I classical receptor, such as glucocorticoid receptor, in the cytoplasm (see Supplementary Information).

To show that MTA1s is necessary for ER cytoplasmic retention in breast cancer cells, we determined the biochemical basis of the MTA1s-mediated retention of ER in the cytoplasm. We created two MTA1s mutants in which we either added a stop codon to delete the unique 33 amino acids (MTA1s^{stop}) or deleted the LRILL motif (MTA1s^{del}). Transient expression of MTA1s^{stop} and MTA1s^{del} did not suppress E2-induced ERE transcription in MCF-7 cells (Fig. 3a). Deletion of LRILL or the 33 amino acids also prevented the interaction of MTA1s with ER in a yeast two-hybrid assay (see Supplementary Information), with endogenous ER in MCF-7 cells (Fig. 3b) and with the AF2 domain of ER (Fig. 3c). To assess the role of LRILL motif in the observed cytoplasmic sequestration of ER, we examined whether deletion of LRILL would release ER to the nucleus in oestradiol-stimulated MCF-7 cells. Expression of MTA1s^{del}, but not wild-type MTA1s, resulted in the presence of ER in the nucleus in cells that expressed MTA1s^{del} (Fig. 3d). Together, these findings indicate that the LRILL motif has a central role in cytoplasmic sequestration of ER by MTA1s.

The observed MTA1s interaction with ER in the cytoplasm impaired the nuclear functions of ER, because ZR75/MTA1s cells showed a significant reduction in both basal and oestradiol-inducible expression of ER-target genes such as pS2 (ref. 11 and Fig. 3e, f), which suggested that overexpression of MTA1s might severely impair the genomic functions of ER in breast cancer cells. Expression of MTA1s enhanced the ability of ZR-75R or MCF-7 cells to grow in an anchorage-independent manner in soft agar, and MTA1s-expressing cells showed a differential response to pure and partial anti-oestrogenic agents (see Supplementary Information). Consistent with more aggressive phenotypes, MTA1s-expressing MCF-7 cells were also more tumorigenic in nude mice in the absence of any oestradiol treatment (see Supplementary Information).

The observed inhibitory effect of ICI 182780 on the anchorage-

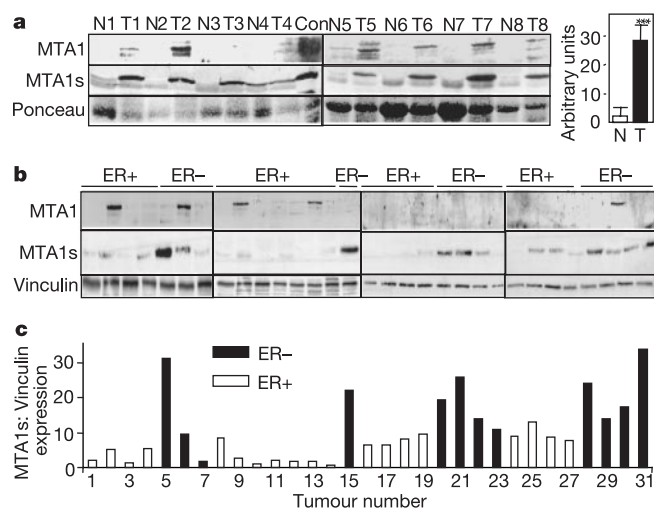
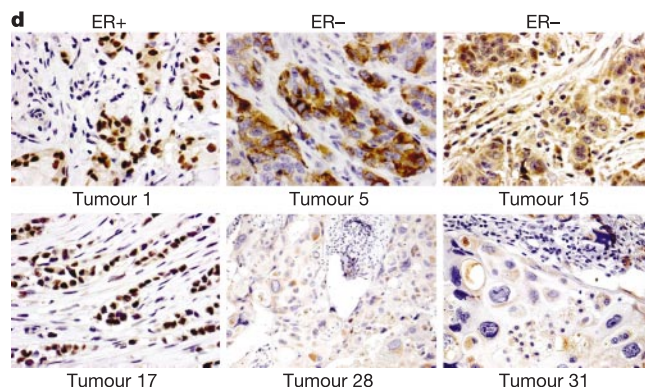


Figure 4 Expression of MTA1s in human breast tumours. **a**, Amounts of MTA1 and MTA1s proteins in paired tumour and adjacent normal tissue from the same individual. N, normal; T, tumour. Con (lane 9) indicates a positive control T7-MTA1s transfected lysate. **b**, Expression of MTA1, MTA1s and vinculin proteins in ER-positive and ER-negative breast tumours. **c**, Quantification of MTA1s expression. **d**, Nuclear and cytoplasmic ER in



tumours with differentially expressing MTA1s. Paraffin-embedded ER-negative tumours (samples 5, 15, 28 and 31) and ER-positive tumours (samples 1 and 17) were immunostained with a monoclonal antibody against ER- α . Original magnification, $\times 2,000$.

independent growth of ZR/MTA1s cells with impaired genomic functions of ER suggested that ICI 182780 might interfere with a non-genomic function of MTA1s-sequestered ER in the cytoplasm. Oestradiol stimulates the p42/44 mitogen-activated kinase (MAPK)^{12,13} pathway and it may be involved in the growth factor receptor-bound protein 2 (GRB2)–son of sevenless (SOS)–Shc pathway¹⁴. We found that oestradiol rapidly stimulates the formation of the GRB2–SOS complex in an ICI-182780-sensitive manner in ZR75/pcDNA cells (Fig. 3g). In contrast to ZR/pcDNA cells, exponentially growing ZR/MTA1s cells contained significant amounts of activated GRB2–SOS complex (Fig. 3h) and MAPK kinase 1 (MEK1; Fig. 3i). In addition, there was significant basal activation of MAPK in ZR/MTA1s cells, which was induced by E2 and blocked by ICI 182780 (Fig. 3j); these results were confirmed by confocal scanning microscopy (see Supplementary Information). There seems to be a close relationship between the amounts of MTA1s and MAPK activation in ER-positive breast cancer cells, because transient expression of MTA1s in ER-negative MDA-MB435 cells failed to activate MAPK (see Supplementary Information). These findings suggest that MTA1s-sequestered ER in the cytoplasm is functional and may contribute towards the non-genomic effects of ER.

To determine the significance of MTA1s in human cancers, we examined the expression of MTA1s in human breast tumours by western blot analysis. Amounts of both MTA1s and MTA1 were higher in tumours than in adjacent normal tissue from the same individual (Fig. 4a). We also examined the amounts of MTA1 and MTA1s in a series of retrospective specimens from ER-positive (tumours with nuclear ER) and ER-negative (tumours without nuclear ER) breast tumours. We found that the quantities of MTA1s, but not MTA1, were significantly higher in the ER-negative tumours than in the ER-positive ones (Fig. 4b). Quantification, and normalization of the MTA1s signal to that of vinculin, indicated a fourfold higher expression of MTA1s in ER-negative tumours than in ER-positive tumours (Fig. 4c). In addition, increased amounts of MTA1s in ER-negative tumours were associated with the presence of cytoplasmic ER in tumour cells (Fig. 4d). Considering that MTA1s sequesters ER in the cytoplasm and that MTA1s expression was high in ER-negative tumours, these findings might explain, in part, the loss of nuclear ER in breast tumours with increased MTA1s.

In summary, we have identified a pathway that may redirect ER signalling by sequestering ER in the cytoplasm, blocking the activation of ER-responsive pathways in the nuclear compartment, and stimulating the non-genomic functions of ER in the cytoplasm. Regulation of the cellular localization of ER by MTA1s represents a previously unknown mechanism for controlling the nuclear functions of ER. □

Methods

Cell cultures and antibodies

Breast cancer cells⁷ were maintained in DMEM:F12 (1:1) medium supplemented with 10% fetal calf serum. ZR-75R cells were supplemented with 1 mM sodium pyruvate. Antibodies against MTA1 (clone A18) and GRB2 were purchased from Santa Cruz and Upstate Biotechnology, respectively. Antibodies against MEK1, phospho-MEK1 (S217/221), phospho-p42/44 MAPK (T202/Y204) were obtained from New England Biolabs, and antibodies against SOS from Signal Transduction. Antibodies against ER- α for confocal immunostaining (clone H184), paraffin immunostaining (clone 1D5-6F11) and protein experiments (clone AER311) were purchased from Santa Cruz, Neomarkers and Upstate Biotechnology, respectively. We crosslinked the 22-amino-acid MTA1s peptide to the keyhole cyanate vector protein and used it to immunize New Zealand white rabbits at Research Genetics. The MTA1s antibody was purified from serum using MTA1s–GST fusion protein, immunoblotted onto nitrocellulose membrane, eluted with MAPS buffer (pH 2.8), neutralized with Tris buffer (pH 8.0), and used at 1:25 dilution.

Cloning and construction of MTA1s mutants

We used a 288-bp MTA1 DNA fragment amplified by RT–PCR from the MCF-7 cells as a probe to isolate the MTA1 cDNAs using conventional molecular biology methods. Mutation of MTA1s was done by the Quick-change kit (Stratagene) using specific primers (see Supplementary Information).

Transfection and promoter–reporter assay

We carried out transfection using a Fugene-6 kit (Roche) according to the manufacturer's instructions. Cells were transiently co-transfected with a reporter construct and β -galactosidase, and assayed as described³. We carried out *in vitro* transcription and translation of the MTA1s protein using the TNT transcription–translation system (Promega).

GST pull-down assay and RT–PCR

The GST pull-down assays were done by incubating GST or GST fusion proteins immobilized on GST beads with *in vitro* translated ³⁵S-labelled protein (see Supplementary Information). RT–PCR was done by the access RT–PCR System (Promega and Ambion, respectively) using specific primers (see Supplementary Information).

Two-hybrid screening

A yeast two-hybrid based assay was carried out with the Matchmaker Yeast Two-Hybrid System-3 (Clontech) in accordance with the manufacturer's protocol (see Supplementary Information).

Immunofluorescence confocal assays

The cellular localization of T7–MTA1s, endogenous MTA1s, ER and phosphorylated MAPK was determined using indirect immunofluorescence as described³ (see Supplementary Information).

Soft-agar growth and tumorigenesis assays

Soft-agar colony growth assays were done as described³. For xenografts studies, 5×10^6 cells were implanted subcutaneously into the mammary fat pad of 10 nude mice per group and allowed to grow for 21 days. We measured tumour size as described¹⁵.

Human tissue samples

Residual tissue specimens from the MD Anderson Breast Cancer Core Laboratory were snap frozen in liquid nitrogen and stored at -80°C (see Supplementary Information).

Received 31 December 2001; accepted 14 May 2002; doi:10.1038/nature00889.

- Toh, Y., Pencil, S. D. & Nicolson, G. L. A novel candidate metastasis-associated gene, mta1, differentially expressed in highly metastatic mammary adenocarcinoma cell lines. cDNA cloning, expression, and protein analyses. *J. Biol. Chem.* **269**, 229–263 (1994).
- Nawa, A. *et al.* Tumor metastasis-associated human MTA1 gene: its deduced protein sequence, localization and association with breast cancer cell proliferation using antisense phosphorothioate oligonucleotides. *J. Cell. Biochem.* **79**, 202–212 (2000).
- Mazumdar, A. *et al.* Transcriptional repression of oestrogen receptor by metastasis-associated protein 1 corepressor. *Nature Cell Biol.* **3**, 30–37 (2001).
- Shapiro, M. B. & Senapathy, P. RNA splices junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acid Res.* **15**, 7155–7174 (1987).
- Crawford, J., Ianzano, L. & Savino, M. The PISLRE gene: structure, exon skipping and exclusion as tumour suppressor in breast cancer. *Genomics* **56**, 90–97 (1999).
- Benz, C. C. *et al.* Estrogen-dependent, taxifen-resistant tumorigenic growth of MCF-7 cells transfected with HER2/neu. *Breast Cancer Res. Treat.* **24**, 85–95 (1992).
- Kumar, R. *et al.* Overexpression of HER2 modulates Bcl-2, Bcl-X_i and Tamoxifen-induced apoptosis in human MCF-7 breast cancer cells. *Clin. Cancer Res.* **2**, 1215–1219 (1996).
- Vadlamudi, R. *et al.* PELP1, a novel human protein that acts as a bifunctional regulator of steroid receptors. *J. Biol. Chem.* **276**, 38272–38279 (2001).
- Henttu, P. M., Kalkhoven, E. & Parker, M. G. AF-2 activity and recruitment of steroid receptor coactivator 1 to the estrogen receptor depend on a lysine residue conserved in nuclear receptors. *Mol. Cell. Biol.* **17**, 1832–1839 (1997).
- Kumar, V. *et al.* Functional domains of the human estrogen receptor. *Cell* **51**, 941–951 (1987).
- Berry, B. *et al.* Estrogen-responsive element of the human pS2 gene is an imperfectly palindromic sequence. *Proc. Natl Acad. Sci USA* **86**, 1218–1222 (1989).
- Migliaccio, A. *et al.* Tyrosine kinase/p21ras/MAP-kinase pathway activation by estradiol-receptor complex in MCF-7 cells. *EMBO J.* **15**, 1292–1300 (1996).
- Catoria, G. *et al.* Non-transcriptional action of oestradiol and oestrogen triggers DNA synthesis. *EMBO J.* **18**, 2500–2510 (1999).
- Song, R.-X. *et al.* Linkage of estrogen action to MAP kinase activation through ER α association with Shc signalling pathway. *Mol. Endocrinol.* **16**, 116–127 (2002).
- Mandal, M. *et al.* Growth factors regulation of heterogeneous nuclear ribonucleoprotein expression and functions. *J. Biol. Chem.* **276**, 9699–9704 (2001).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements

We thank B. S. Katzenellenbogen and M. R. Parker for AF2 fusion protein; D. McDonnell for 3X-ERE TATA; D. Shapiro for ERE-reporter systems; D. Nguyen for cloning; and W. Schmid for the GRE-CAT promoter constructs. Susan G. Komen Breast Cancer Foundation and National Institutes of Health grants (to R.K.) supported this study.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for material should be addressed to R.K. (e-mail: rkumar@mdanderson.org).

Biomimetic synthesis and optimization of cyclic peptide antibiotics

Rahul M. Kohli^{*†}, Christopher T. Walsh^{*} & Michael D. Burkart^{*†‡}

^{*} Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA

[‡] Present address: Department of Chemistry and Biochemistry, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA

[†] These authors contributed equally to this work

Molecules in nature are often brought to a bioactive conformation by ring formation (macrocyclization)¹. A recurrent theme in the enzymatic synthesis of macrocyclic compounds by non-ribosomal and polyketide synthetases is the tethering of activated linear intermediates through thioester linkages to carrier proteins, in a natural analogy to solid-phase synthesis². A terminal thioesterase domain of the synthetase catalyses release from the tether and cyclization^{3,4}. Here we show that an isolated thioesterase can catalyse the cyclization of linear peptides immobilized on a solid-phase support modified with a biomimetic linker, offering the possibility of merging natural-product biosynthesis with combinatorial solid-phase chemistry. Starting from the cyclic decapeptide antibiotic tyrocidine A, this chemo-enzymatic approach allows us to diversify the linear peptide both to probe the enzymology of the macrocyclizing enzyme, TycC thioesterase, and to create a library of cyclic peptide antibiotic products. We have used this method to reveal natural-product analogues of potential therapeutic utility; these compounds have an increased preference for bacterial over eukaryotic membranes and an improved spectrum of activity against some common bacterial pathogens.

Many pharmacologically important natural products are constrained by macrocyclization, from the cyclic non-ribosomal peptides (NRP) tyrocidine A and cyclosporin, to the polyketide (PK) antibiotic erythromycin, and the hybrid peptide/polyketide drugs

rapamycin and epothilone. These macrocyclic molecules are constructed in nature by similar logic using NRP synthetases (NRPS), PK synthetases or hybrid NRP/PK synthetases^{2,3}. As seen in the biosynthesis of the cyclic decapeptide antibiotic tyrocidine A by an NRP synthetase (Fig. 1), the linear molecule is first constructed by a multi-domain, modular enzymatic 'assembly line' with bio-synthetic intermediates tethered to successive modules, as in solid-phase synthesis⁵. The final step in enzymatic macrocycle synthesis is the cyclization of a tethered linear precursor by an integrated carboxy-terminal thioesterase (TE) domain. We have shown that TycC TE, the isolated C-terminal TE domain excised from the larger synthetase protein, can catalyse cyclization of a soluble peptide thioester that serves as a surrogate for the protein-bound substrate⁶. Systematic alteration in the soluble substrate has shown that the TE domain is a permissive cyclization catalyst, tolerating variations in the length of the peptide substrate and the identity of most individual side chains^{6,7}. Given the lack of a general chemical solution to macrocyclization, the liberal cyclization abilities of TycC TE offer a prospect for exploring diverse macrocyclic structures via chemoenzymatic synthesis.

Building on the analogy of NRP synthesis to solid-phase peptide synthesis (SPPS), we chose to adopt the synthesis of linear substrates for the TE to a solid-phase amine-terminal polyethylene glycol amide (PEGA) resin (Fig. 1). In a biomimetic manner, a resin-bound substrate could substitute for the natural protein-bound substrate, and the isolated TE domain could catalyse release from the resin and cyclization. We anticipated that diverse peptides could then be constructed using solid-phase synthetic strategies to achieve two aims: (1) to probe the enzymology of biosynthesis of macrocyclic molecules, and (2) to construct natural-product analogues to understand or optimize their biological activity. To achieve these goals, the solid-phase linker was designed to satisfy two anticipated requirements: stability to peptide synthesis and recognition of the resin-bound substrate by the cyclizing enzyme. As oxo-ester bonds are stable to peptide synthesis based on Fmoc (9-fluorenylmethoxycarbonyl), we synthesized and kinetically characterized **4**, a soluble decapeptide *N*-acetyethanolamine ester version of the linear tyrocidine A substrate (Fig. 1). The observed modest decrease in catalytic rate constant (k_{cat}) of cyclization, despite the significantly decreased thermodynamic activation, established the viabi-

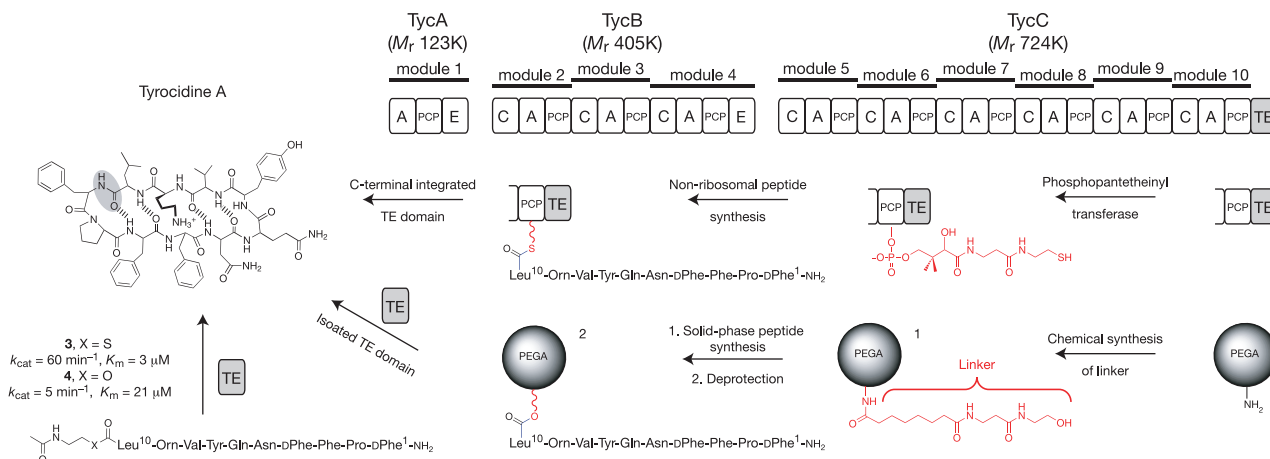


Figure 1 Natural versus biomimetic macrocycle synthesis. The enzymatic assembly line involved in biosynthesis of the cyclic cationic antimicrobial peptide, tyrocidine A, consists of three large synthetase proteins with ten modules. A carrier protein (PCP) in each module is loaded with a phosphopantetheine prosthetic group (red). The individual modules contain domains (A, adenylation; C, condensation; E, epimerization) that load amino-acid building blocks onto the thioester tether and condense via successive peptide bond forming reactions giving the terminal PCP domain loaded with a linear decapeptide.

The terminal thioesterase domain (TE) catalyses head-to-tail cyclization. In the biomimetic synthetic strategy, a linker (red) that mimics phosphopantetheine is chemically synthesized onto a solid-phase resin. Solid-phase peptide synthesis is used to construct **2**, a tethered linear peptide, which can then serve as a substrate for cyclization by the TE domain excised from the synthetase proteins. The excised TE domain can catalyse cyclization of soluble peptidyl thioester (**3**) and ester (**4**) substrates with listed kinetic parameters (k_{cat} and K_m (Michaelis constant)). M_r , relative molecular mass.

lity of an oxo-substitution for the natural thioester phosphopantetheine linkage. To allow for recognition of the substrate by the TE enzyme, we designed a solid-phase linker structurally homologous to the natural phosphopantetheine tether. A peptidic linkage consisting of a suberic acid spacer followed by β -alanine and ethanolamine was chosen.

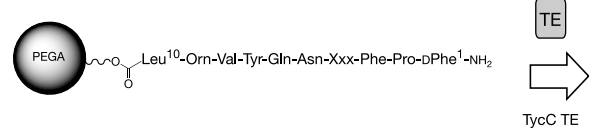
A linear peptide with the sequence of tyrocidine A was synthesized on the primary hydroxyl of **1** by SPPS, and the resin was incubated with purified TycC TE. The enzyme recognized the tethered substrate, and catalysed both release from the solid-phase and cyclization to form tyrocidine A (Fig. 2b). The cyclic product was observed in a ratio of 7:1 to the released linear decapeptide, approximating the ratio observed with soluble substrates. Under the reaction conditions, no non-enzymatic cyclization and minor non-enzymatic hydrolysis were observed.

Tyrocidine A has an amphipathic β -sheet structure, with one face of the β -sheet defined by a single basic ornithine (Orn) residue, and the other face defined by several hydrophobic side chains⁸. Pro-

duced by *Bacillus brevis*, it is a member of the family of cationic peptide antimicrobials, found across the spectrum of phylogeny, including the mammalian disulphide-constrained defensins essential in first-line immune response by neutrophils^{9,10}. We exploited the use of SPPS to target position D-Phe4 in the sequence of tyrocidine A, located in the middle of one strand of the β -sheet. A library of peptides with natural and non-natural amino-acid substituents in the D-Phe4 position was constructed by parallel synthesis in a 96-well format. After incubation of the solid-phase substrates with TycC TE, the filtrate was analysed by high-pressure liquid chromatography coupled to a mass spectrometer (LC/MS), which allowed for both identification and quantification of products. Eight samples coupled in parallel with the natural D-Phe residue demonstrated the reproducibility of the substrate assay (Table 1, entry 7).

The enzymatic mechanism of cyclization involves transacylation of the peptide to an active-site serine, followed by deacylation upon intramolecular attack by the amino-terminal nucleophile¹¹. If the

Table 1 Examination of the D-Phe4 position by biomimetic synthesis of a library of variants

							
Amino acid (XXX)	Cycle (μ M)	Cyc/hyd	Antibiotic	Amino acid (XXX)	Cycle (μ M)	Cyc/hyd	Antibiotic
8-A-3, 6-DO	53.8	11.4		Homophe, L	2.3	0.5	
Ala, D	25.8	9.2	+	2-Furylala, L	6.8	0.5	
Ser, D	48.2	8.5	+	Leu, L	5.1	0.5	
Gln, D	75.8	7.8	+	Chxala, L	1.2	0.4	
Ahx	48.5	7.7		Homopro, D*	4.1	0.4	
Trp, D	82.1	7.7	+	Ser, L	7.0	0.4	
Phe, D	33.1 $\pm$ 2.6	7.1	+	Norval, L	3.9	0.3	
Allo-Thr, D	28.2	7.0	+	Phe, L	3.2	0.3	
Thr, D	43.1	6.2	+	Ala, L	10.4	0.3	
Arg, D	67.6	6.0	+	2-Thienylala, L	2.6	0.3	
Isn	54.4	6.0		Val, L	5.4	0.3	
p-F-Phe, D	24.7	5.9	+	t-Butylgly, L	5.3	0.3	
Styrylala, D	31.2	5.8	+	Ile, L	2.2	0.3	
2-Thienylala, D	30.3	5.7	+	Homoser, L	12.4	0.3	
p-Me-Phe, D	29.2	5.6	+	Abu, L	5.7	0.2	
Chxala, D	16.9	5.5	+	Homocys, L	1.3	0.2	
Val, D	19.1	5.3	+	4-Pyridylala, L	6.0	0.2	
Tyr, D	40.7	5.2	+	Tic, D*	3.9	0.2	
4-Pyridylala, D	38.6	5.1	+	Ornithine, L†	13.6	0.2	+
2-Furylala, D	23.5	4.8	+	Trp, L	9.5	0.2	
2-Napthylala, D	68.4	4.6	+	His, L	6.1	0.2	
Lys, D	90.2	4.3	+	Thr, L	3.6	0.2	
Glu, D†	41.4	4.3		Gln, L	5.9	0.2	
t-Butylgly, D	13.6	4.1	+	Citrulline, L	5.5	0.2	
Ile, D	9.3	4.1	+	Dpr, L†	2.8	0.2	+
Asn, D	49.5	3.6	+	4-Hydroxypro, L	4.7	0.1	
His, D	45.2	3.4	+	Lys, L†	10.9	0.1	+
3,3-Diphe, D	29.0	3.3	+	Dab, L	11.1	0.1	
Ornithine, D	79.1	3.1	+	Arg, L†	4.9	0.1	+
Phenylgly, L	10.3	2.9	+	Aib, L	2.2	0.1	
Phenylgly, D	13.1	2.7	+	No amino acid	2.9	0.1	
Asp, D†	48.1	2.6		Homopro, L	1.9	0.1	
Met, D	7.2	2.5		Phospho-Tyr, L	1.3	0.1	
p-Bz-Phe, D	18.6	2.4	+	Pro, L	3.3	0.1	
Homophe, D	8.0	2.4	+	Glu, L	4.1	0.1	
p-tBu-Phe, D	8.2	2.2		Asp, L	4.6	0.1	
Gly	26.5	2.0		p-NO ₂ -Phe, L	0.3	0.1	
Leu, D	5.9	1.7	+	p-Bz-Phe, L	0.3	0.1	
Asn, L	25.2	1.4		Aad, L	1.3	0.0	
2-Napthylala, L	10.7	1.1		Tic, L	0.5	0.0	
N-Me-Phe, D*	18.2	0.7	+	Diphe, L	0.0	0.0	
Met, L	4.8	0.6		3,5-Diiodo-Tyr, L	0.0	0.0	
Tyr, L	17.3	0.6		Pro, D*	0.0	0.0	
Norleu, L	2.9	0.5		Styrylala, L	0.0	0.0	

Addition of the TE catalyses formation of the cyclization product (shown schematically as β -sheet) or the hydrolysis product. The concentration of cyclic product (in μ M) generated in the reaction is listed, with the data shown ranked in order of decreasing ratio of cyclization to hydrolysis. The products of the reaction were assayed for antibiotic activity against *Bacillus subtilis*. Activity is noted as positive if growth is inhibited at the concentration of cyclic peptide present in the reaction products. 8-A-3, 6-DO, 8-amino-3,6-dioxanoic acid; Ahx, ϵ -aminohexanoic acid; Isn, isonipecotic acid; Abu, 2-aminobutyric acid; Tic, tetrahydroisoquinoline-3-carboxylic acid; Dpr, 2,3-diaminopropionic acid; Dab, 2,4-diaminobutanoic acid; Aib, 2-aminoisobutyric acid; Aad, 2-aminoadipic acid; Chxala, cyclohexylalanine.

*Notable exceptions to the enzymatic preference for D- over L-amino acids.

†Notably inactive D-amino-acid analogues.

‡Notably active L-amino-acid analogues.

substrate is not suitable for cyclization, hydrolysis of intermediate generates linear peptide acid. The cyclization to hydrolysis ratio of the members of the library therefore gave us insight into the enzymology of the TE (Table 1). We discovered a preference for D- over L-amino acids at position four, suggesting steric interactions interfering with formation of the reactive cyclizing conformation. D-amino acids substituted on the backbone amine, such as D-Pro, which are disfavoured from participating in β -sheets, were notable exceptions to the stereochemical trend, suggesting the mechanistic importance of the potential to form a product-like β -sheet near the cyclization junction. Additionally, peptides containing non- α -amino acids were good substrates for cyclization, demonstrating the utility of TE-mediated cyclization for generation of non-natural macrocycles.

The antibiotic activity of tyrocidine A is speculated to arise as a consequence of its amphipathic character^{9,12}. Assay of the antibiotic activity of the reaction products against *Bacillus subtilis* was revealing about both faces of the β -sheet (Table 1), as the side chain of D-amino acids would be oriented towards the hydrophobic face of the peptide, while L-amino acids contribute to the basic face⁸. As the linear tyrocidine A decapeptide demonstrated no antibiotic activity at 1 mM, activity in enzymatic products was attributed to the cyclic products. All peptides containing positively charged L-amino acids at the variable position were potent antibiotics, showing activity despite the low levels of cyclic product formation. The cyclic peptide gramicidin S is a natural analogue to these synthetic tyrocidine A analogues, as it contains an L-Orn residue oriented towards the basic face of the amphipathic antibiotic at the D-Phe 4 position^{12,13}. With D-amino-acid residues, the peptides containing negatively charged D-Glu and D-Asp showed no antibiotic activity. This comprehensive survey of substituents at position four in tyrocidine A points to the importance of a basic face and a non-negative, largely hydrophobic face for potency of tyrocidine A analogues as antibiotics.

Our previous biochemical studies revealed the importance of the N-terminal D-Phe 1 nucleophile to TE-mediated cyclization, as this position was intolerant to substitution by L-Phe or D-Ala with soluble substrates⁶. We now were able to apply our library of amino acids to examine extensively the substrate specificity at position one. Only aromatic D-amino acids, including non-natural heterocycles and aromatics such as D-2-furylalanine and D-2-naphthylalanine,

had cyclization to hydrolysis ratios greater than one, suggestive of a D-specific aromatic binding pocket for the peptide N terminus.

Rather than interacting with specific protein or RNA molecular targets, most cationic peptides are believed to exert their antibiotic effect by non-specific insertion into cell membranes, disrupting osmotic and ionic regulation¹⁴. Their rapid membranolytic effect, as well as the difficulty in selecting resistant mutants *in vitro*, make them appealing pharmacological candidates in the age of growing resistance to conventional antibiotics¹⁵. However, non-specific interaction with membranes has to date limited the utility of tyrocidine A owing to interaction with eukaryotic cells. We next applied biomimetic macrocycle synthesis to find a potentially useful analogue of tyrocidine A with improved selectivity. We selected 8 cyclizing variants at the D-Phe1 position and 24 at the D-Phe4 position that cyclized well and showed antibiotic activity. An 8 $\times$ 24 matrix of cyclic peptides was generated by biomimetic synthesis, with all linear peptides acting as better substrates for cyclization than for hydrolysis. To examine selectivity, we determined both the minimal concentration for antimicrobial activity against *B. subtilis* (MIC) and the minimal concentration for haemolysis of human erythrocytes (MHC)^{16,17}. Most analogues in the matrix had an improved therapeutic profile relative to the naturally occurring antibiotic tyrocidine A, as measured by the therapeutic index (Fig. 2a). In particular, substitution of position 4 by a positive D-amino acid (D-Arg, D-Lys or D-Orn) resulted in the best therapeutic indices. We speculate that the selectivity of these dibasic analogues is due to a preferred interaction with the acidic outer-leaflet phospholipids of bacterial cell membranes over zwitterionic phospholipids in eukaryotic cell membranes¹⁴.

Two of the best analogues in our assay, Δ A02 and Δ F10, along with the authentic tyrocidine A sequence, were selected for preparative-scale chemoenzymatic synthesis. When assayed against a series of pathogenic bacteria, the dibasic analogues Δ A02 and Δ F10 showed broad-spectrum activity against Gram-positive and Gram-negative organisms, and improved therapeutic profiles relative to the natural product (Table 2). The characteristics of the cationic dibasic peptide analogues of tyrocidine A make them appealing candidates for further evaluation for antimicrobial therapy, may generate new structure/activity rules for cyclic antimicrobial peptides, and establish the potential of TE domains in combinatorial biosynthesis of bioactive molecules.

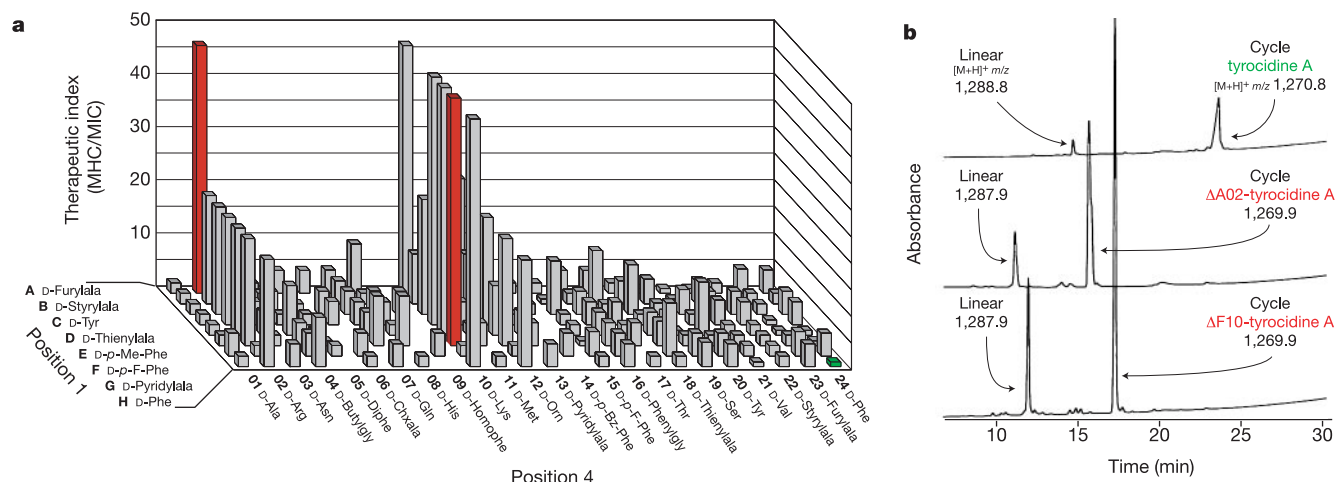


Figure 2 Improved analogues of tyrocidine A. **a**, Therapeutic indices of the analogues synthesized by the biomimetic method with variants at the D-Phe 1 position (**A–H**) and at the D-Phe 4 position (**01–24**). The analogues with positively charged D-amino acids in place of D-Phe 4 show the greatest increase in selectivity for bacterial membranes. The

authentic antibiotic tyrocidine A is highlighted in green. The two analogues selected for further characterization are shown in red. **b**, LC/MS trace of reaction products released by Tyc TE enzymatic action on solid-phase resin with natural tyrocidine A and variants Δ A02 and Δ F10.

Table 2 Improved specificity and spectrum of activity of tyrocidine A analogues

	Tyrocidine A	ΔA02-tyrocidine A	ΔF10-tyrocidine A
Eukaryotic (MHC, μM)			
Human erythrocytes	4	60	150
Gram negative (MIC, μM)			
<i>Escherichia coli</i>	>290	120	290
<i>Pseudomonas aeruginosa</i>	290	30	37
<i>Pseudomonas putida</i>	>290	30	74
Gram positive (MIC, μM)			
<i>Bacillus subtilis</i>	1	2	5
MRSA	2	7	18
<i>Staphylococcus epidermidis</i>	2	4	9

MHC, minimum haemolytic concentration; MIC, minimum inhibitory concentration; MRSA, methicillin-resistant *S. aureus*.

In the natural biosynthesis of many macrocyclic compounds, a synthetase constructs a linear molecule tethered via a phosphopantetheine linker to a carrier protein followed by enzymatic release and cyclization. In our chemoenzymatic strategy, we mimic this process, and use the strengths of solid-phase synthesis to access more than 300 linear substrates tethered to a solid-phase resin for cyclization by a permissive enzyme catalyst, an NRPS-derived TE domain. The diversity could be further augmented by variation at other positions to establish fully the sequence space requirements of this TE, or by use of TE domains from other synthetases that catalyse diverse cyclization regiochemistries⁴. In contrast to chemical and intein-mediated cyclization, this method exploits enzymatic specificity, generating a regiospecific, monomeric product, while also incorporating non-natural and non-proteinogenic amino acids^{17–19}. As we can explore the sequence space about a natural product, an optimized analogue, once identified, could be synthesized by re-engineering the NRPS of the biological producer^{2,20}. More generally, the use of the biomimetic solid-phase resin as a mimic of a phosphopantetheine-containing carrier protein can now be effectively used to probe the diverse natural structures made by NRP, PK, hybrid NRP/PK and fatty-acid synthetases, as well as the individual domains and modifying enzymes used in these biosynthetic pathways. □

Methods

Linker synthesis

PEGA resin (10 g) was coupled to 3.6 ml (5 equiv.) monomethyl suberic acid by the PyBOP coupling procedure (adding 4.9 equiv. benzotriazol-1-yloxytripyrrolidinophosphonium hexafluorophosphate (PyBOP), 5 equiv. 1-hydroxybenzotriazole (HOBt), and 10 equiv. diisopropylethylamine (DIPEA) in dimethylformamide (DMF), followed by agitation for 2 h. The resin is washed with DMF and coupling repeated). Base deprotection (incubation in THF/MeOH/10N NaOH (6/3/1) to cover beads, agitation for 30 min, followed by acid wash) yields the free acid. 3 g (5 equiv.) β-alanine methyl ester•HCl were coupled to the resin by the PyBOP coupling procedure followed by base deprotection. 7.8 g (20 equiv.) ethanolamine•HCl was coupled to the resin by the PyBOP coupling procedure generating 1.

Synthesis of peptide substrate libraries and enzymatic reactions

Automated SPPS was performed on a PerSeptive Biosystems 9050 automated peptide synthesizer. The first amino acid was coupled to 1 by 5 equiv. 1-(mesitylene-2-sulphonyl)-3-nitro-1H-1,2,4-triazole (MSNT) and 3.75 equiv. 1-methylimidazole in DMF. For library synthesis, residues 10–5 of tyrocidine A were synthesized by automated SPPS, and the resin was distributed with 2 μmol per well in a 96-well, 2-ml polypropylene filter plate fitted with a combinatorial clamp. Subsequent amino acids were coupled following standard Fmoc peptide synthesis techniques, and the decapeptides deprotected by trifluoroacetic acid/water/triisopropylsilane (95/2.5/2.5). Excised Tyc TE domain was overexpressed and purified as previously reported⁶. Enzymatic catalysis on resin-bound substrates were performed by incubating pre-swollen resin with 10 μM Tyc TE in 800 μl of 50 mM MOPS, pH 7.0 at 24 °C for 2 h. Products were eluted from the filter plates by centrifugation, and the resin was washed with 200 μl CH₃CN. Pooled elutions and washes were analysed by a Shimadzu QP8000 electrospray ionization LC/MS to identify linear and cyclic products. The determined extinction coefficient for the linear tyrocidine peptide at 220 nm wavelength was used to quantify product formation. Synthesis and enzymatic cyclization of linear tyrocidine-*N*-acetylthanolamine (4) was performed according to linear tyrocidine-*N*-acetylcysteamine (3), replacing *N*-acetylcysteamine with *N*-acetylthanolamine⁶.

Determination of bioactivity

Serial 2-fold dilutions of peptide were prepared in microtitre plates. The peptide solutions were dried under vacuum. For MIC determination, 80 μl of 1/10,000 dilution of overnight culture of bacteria in LB media was added to each well. The plates were incubated at 30 °C for 16–20 h, and MIC was determined by visual assay as the lowest concentration resulting in complete inhibition of growth. Bacteria used were *B. subtilis* PY79, *Pseudomonas aeruginosa* PA01, *Escherichia coli* clinical isolate, *Pseudomonas putida* WCS358 and *Staphylococcus epidermidis* R94-10. Assays with methicillin-resistant *Staphylococcus aureus* (MRSA; ATCC 33591) were carried out by MDS PanLabs Pharmacology Services. For determination of MHC, 10 ml freshly collected human blood with buffered citrate was layered over 4 ml histopaque-1077 (Sigma). After centrifugation, the bottom layer containing packed erythrocytes was preserved. The cells were diluted with phosphate-buffered saline to 1/100 packed volume of cells, and 100 μl was added to each well. Plates were incubated with rocking at 24 °C for 16 h, and the MHC was determined visually by loss of turbidity^{16,17}.

Preparative-scale synthesis

Scale-up was carried out by synthesis of target sequences on 100 μmol resin 1 as described above. The resin was incubated with 4 ml of 180 μM Tyc TE in 50 mM MOPS, pH 7.0. After 30 min, filtrate was collected and reaction was repeated. After second collection, resin was washed with 2 ml CH₃CN and filtrate pooled, resulting in aggregation of Tyc TE and the cyclic peptides. The reactions were centrifuged and supernatant containing linear peptides discarded. The pellet was redissolved in 1:7 trifluoroethanol:water and filtered, affording the purified cyclic peptides for quantification and assay.

Received 8 April; accepted 6 June 2002; doi:10.1038/nature00907.

- Rizo, J. & Gierasch, L. M. Constrained peptides: models of bioactive peptides and protein substructures. *Annu. Rev. Biochem.* **61**, 387–418 (1992).
- Cane, D. E., Walsh, C. T. & Khosla, C. Harnessing the biosynthetic code: combinations, permutations, and mutations. *Science* **282**, 63–68 (1998).
- Marahiel, M. A., Stachelhaus, T. & Mootz, H. D. Modular peptide synthetases involved in nonribosomal peptide synthesis. *Chem. Rev.* **97**, 2651–2674 (1997).
- Keating, T. A. *et al.* Chain termination steps in nonribosomal peptide synthetase assembly lines: directed acyl-S-enzyme breakdown in antibiotic and siderophore biosynthesis. *ChemBioChem* **2**, 99–107 (2001).
- Lambolot, R. H. *et al.* A new enzyme superfamily—the phosphopantetheinyl transferases. *Chem. Biol.* **3**, 923–936 (1996).
- Trauger, J. W., Kohli, R. M., Mootz, H. D., Marahiel, M. A. & Walsh, C. T. Peptide cyclization catalyzed by the thioesterase domain of tyrocidine synthetase. *Nature* **407**, 215–218 (2000).
- Kohli, R. M., Trauger, J. W., Schwarzer, D., Marahiel, M. A. & Walsh, C. T. Generality of peptide cyclization catalyzed by isolated thioesterase domains of nonribosomal peptide synthetases. *Biochemistry* **40**, 7099–7108 (2001).
- Kuo, M. C. & Gibbons, W. A. Nuclear Overhauser effect and cross-relaxation rate determinations of dihedral and transannular interproton distances in the decapeptide tyrocidine A. *Biophys. J.* **32**, 807–836 (1980).
- Zaslhoff, M. Antimicrobial peptides of multicellular organisms. *Nature* **415**, 389–395 (2002).
- Nizet, V. *et al.* Innate antimicrobial peptide protects the skin from invasive bacterial infection. *Nature* **414**, 454–457 (2001).
- Li, J., Szittner, R., Derewenda, Z. S. & Meighen, E. A. Conversion of serine-114 to cysteine-114 and the role of the active site nucleophile in acyl transfer by myristoyl-ACP thioesterase from *Vibrio harveyi*. *Biochemistry* **35**, 9967–9973 (1996).
- Izumiya, N. *Synthetic Aspects of Biologically Active Cyclic Peptides: Gramicidin S and Tyrocidines* (Wiley, New York, 1979).
- Hull, S. E., Karlsson, R., Main, P., Woolfson, M. M. & Dodson, E. J. The crystal structure of a hydrated gramicidin S-urea complex. *Nature* **275**, 206–207 (1978).
- Matsuzaki, K. Why and how are peptide-lipid interactions utilized for self-defense? *Biochim. Biophys. Acta* **1462**, 1–10 (1999).
- Hancock, R. E. & Lehrer, R. Cationic peptides: a new source of antibiotics. *Trends Biotechnol.* **16**, 82–88 (1998).
- Young, J. D., Leong, L. G., DiNome, M. A. & Cohn, Z. A. A semiautomated hemolysis microassay for membrane lytic proteins. *Anal. Biochem.* **154**, 649–654 (1986).
- Kondejewski, L. H. *et al.* Dissociation of antimicrobial and hemolytic activities in cyclic peptide diastereomers by systematic alterations in amphipathicity. *J. Biol. Chem.* **274**, 13181–13192 (1999).
- Scott, C. P., Abel-Santos, E., Jones, A. D. & Benkovic, S. J. Structural requirements for the biosynthesis of backbone cyclic peptide libraries. *Chem. Biol.* **8**, 801–815 (2001).
- Ovchinnikov, Y. A. & Ivanov, V. T. in *The Proteins* Vol. 5 (eds Neurath, H. & Hill, R. L.) 391–399 (Academic, New York, 1982).
- Stachelhaus, T., Schneider, A. & Marahiel, M. A. Rational design of peptide antibiotics by targeted replacement of bacterial and fungal domains. *Science* **269**, 69–72 (1995).

Acknowledgements

We thank R. Kolter for providing strains for antibacterial assays, and the Harvard Center for Proteomics for access to equipment. We also thank M. D. Burke and J.-M. Gauguier for discussions. This work was supported by the NIH (C.T.W.). R.M.K. is supported by the Medical Scientist Training Program, and M.D.B. was an NIH post-doctoral fellow.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.T.W. (e-mail: christopher_walsh@hms.harvard.edu).

Three-dimensional structure of the bacterial protein-translocation complex SecYEG

Cécile Breyton^{*†}, Winfried Haase^{*}, Tom A. Rapoport[‡],
Werner Kühlbrandt^{*} & Ian Collinson^{*‡}

^{*} Max-Planck-Institut für Biophysik, Abteilung Strukturbiologie, Heinrich-Hoffmann-Straße 7, 60528 Frankfurt am Main, Germany

[‡] Department of Cell Biology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA

[†] Permanent address: CNRS-UMR7099, IBPC, 13 rue P. et M. Curie, 75005 Paris, France

Transport and membrane integration of polypeptides is carried out by specific protein complexes in the membranes of all living cells. The Sec transport path provides an essential and ubiquitous route for protein translocation¹. In the bacterial cytoplasmic membrane, the channel is formed by oligomers of a heterotrimeric membrane protein complex consisting of subunits SecY, SecE and SecG^{2,3}. In the endoplasmic reticulum membrane, the channel is formed from the related Sec61 complex⁴. Here we report the structure of the *Escherichia coli* SecYEG assembly at an in-plane resolution of 8 Å. The three-dimensional map, calculated from two-dimensional SecYEG crystals, reveals a sandwich of two membranes interacting through the extensive cytoplasmic domains. Each membrane is composed of dimers of SecYEG. The monomeric complex contains 15 transmembrane helices. In the centre of the dimer we observe a 16 × 25 Å cavity closed on the periplasmic side by two highly tilted transmembrane helices. This may represent the closed state of the protein-conducting channel.

The fundamental mechanism underlying protein translocation through membranes cannot be understood without detailed information of the three-dimensional (3D) structure of the membrane proteins that form the translocation channel. So far, structural information has been limited to electron microscopy and single-particle analysis of the eukaryotic Sec61 or bacterial SecYEG complexes at low resolution^{5–8}. To obtain a higher-resolution structure, crystalline membrane vesicles of the *E. coli* SecYEG were grown⁹. These crystals form by incorporation of the purified, detergent-solubilized complex into a phospholipid bilayer by slow removal of the detergent in the presence of lipids. This method has yielded SecYEG active in protein translocation^{2,3,9}. The 3D structure of SecYEG was determined by the analysis of electron micrographs of crystals, tilted at angles of 0–55°. Amplitudes and phases of structure factors extracted from the images were combined to calculate a 3D map of SecYEG at 8 Å resolution in the membrane plane (Table 1, Fig. 1).

A side view of the map shows an unusual feature of the SecYEG crystals: they are composed of two membrane layers related to one another by a two-fold screw (2₁) axis (Fig. 1a). Therefore, the projection map⁹ was a superposition of two crystalline membranes. This explains why it did not reveal the boundaries of the complex or its oligomeric state. The two membranes are each about 40 Å thick and separated by a zone of lower density roughly 25 Å wide. As each membrane layer contains only the noncrystallographic two-fold symmetry, the complexes within one membrane have the same orientation, and thus, owing to the crystallographic 2₁ axis, the two outer surfaces of the sandwich crystal are the same. The orientation of the protein in the crystals was investigated by immuno-electron microscopy, employing antibodies directed against the carboxy-terminal region of SecG or SecY, which are known to be located on the periplasmic and cytoplasmic sides of the membrane, respectively^{10,11} (Fig. 2a, b). The SecG antibodies labelled the crystals, whereas the SecY antibodies did not (Fig. 2c, e). Thin sections of the crystals revealed that the SecG antibodies bound to both sides

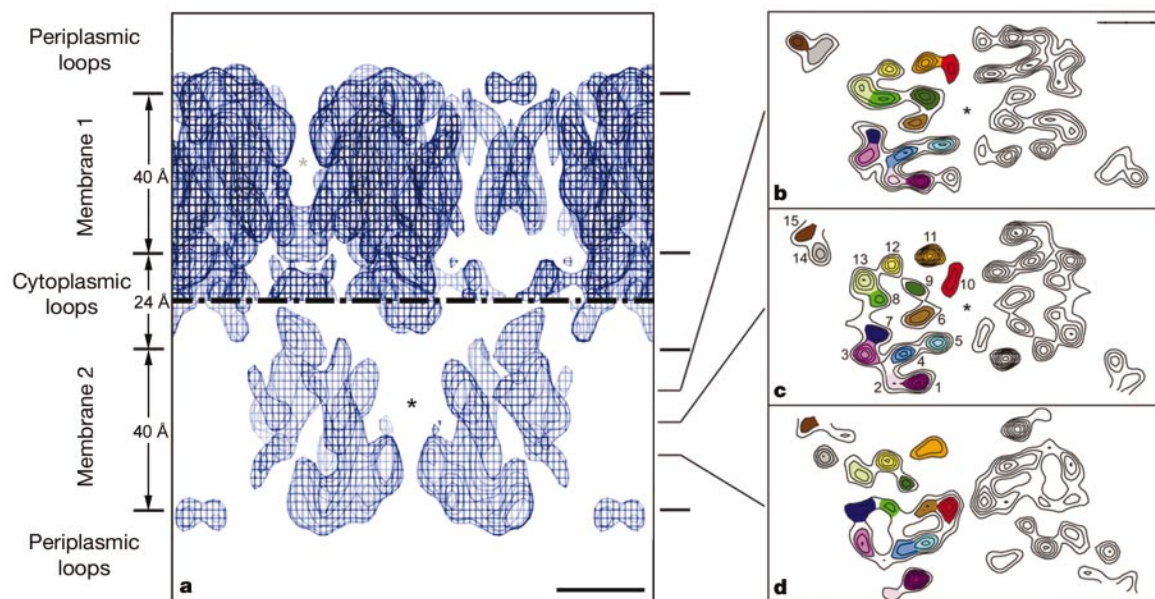


Figure 1 Double membrane crystal of SecYEG. **a**, Side view of the map of SecYEG at 8 Å in-plane resolution, with applied noncrystallographic two-fold symmetry, contoured at 1.35 s.d. A region of lower density, containing the cytoplasmic loops from both layers, separates the two membranes. The membranes are related to one another by the crystallographic 2₁ axis. The upper membrane is a slab one-unit cell deep. The lower membrane is a thinner slab centred on the two-fold axis of the dimer, revealing a putative

channel (asterisk), which is closed on the periplasmic side. **b–d**, Horizontal sections, 8 Å apart, through the SecYEG dimer, as indicated in **a**. In each section, 15 peaks of density, corresponding to sections of membrane-spanning helices, are resolved within a monomer of SecYEG. They are arbitrarily numbered from 1 to 15 and colour coded. Scale bar, 20 Å.

of the double membrane (Fig. 2d). Therefore, the outer surfaces of the crystal correspond to the periplasmic side of the protein, and the more extensive cytoplasmic domains provide the crystal contacts between the two membranes (Fig. 1a). These features are consistent with topology models indicating that twice as many residues of SecYEG are exposed on the cytoplasmic surface than on the periplasmic side^{10,11}.

Within a single membrane, the SecYEG complex forms dimers that are clearly separated from one another by featureless, lipid-containing regions (Fig. 3a). The two monomers within the dimer are related to each other by noncrystallographic two-fold symmetry. In each SecYEG monomer, 15 rod-shaped densities—characteristic of membrane-spanning α -helices—are resolved (arbitrarily labelled 1–15; Figs 1b–d and 3), consistent with predictions based on polypeptide hydropathy. Helices 1, 3 and 12 run roughly perpendicular to the membrane plane; helices 2, 4, 6, 11 and 13–15 have tilt angles of 5–20° relative to the membrane normal, whereas helices 5 and 7–10 are highly tilted by 30–45° (Fig. 3a). All membrane-spanning helices are well resolved at a contour level of one s.d. (Figs 1b–d and 3).

The monomer accommodates 13 tightly packed helices, with centre-to-centre distances of 7–13 Å. The two remaining helices are isolated from the main bundle. SecY, SecE and SecG are predicted to contain 10, 3 and 2 membrane-spanning α -helices, respectively. Because SecG is not required for translocation *in vitro*^{2,12} or *in vivo*¹³, a peripheral location of this subunit in the complex seems likely.

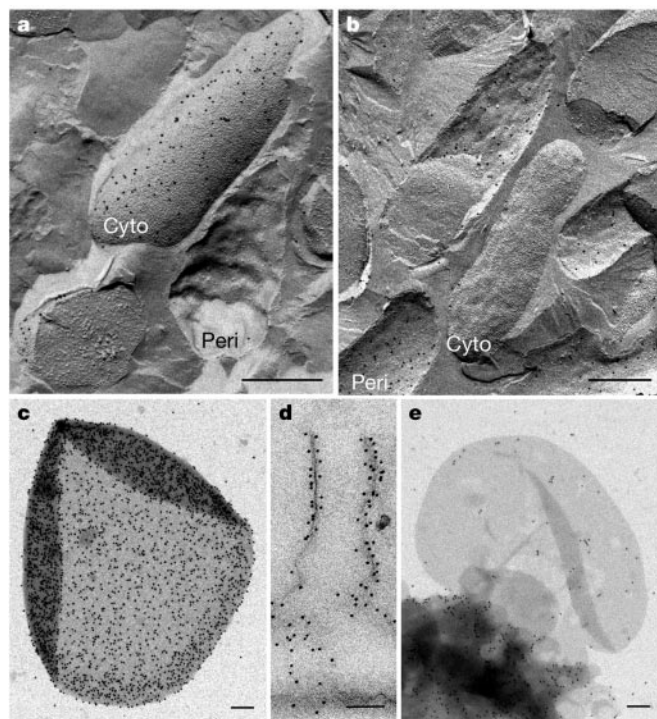


Figure 2 Immunogold labelling with antibodies directed against C-terminal fragments of SecG (**b–d**) or of SecY (**a** and **e**). *E. coli* cells overexpressing the SecYEG complex were freeze-fractured, splitting the inner membrane into convex (cytoplasmic, cyto) and concave (periplasmic, peri) leaflets. Immunogold labelling was performed as described²⁴. As expected, the SecG antibodies labelled only the periplasmic surface (**b**), whereas the SecY antibodies bound only to the cytoplasmic surface of the bacterial membrane (**a**). SecYEG crystals were immunogold labelled on the electron microscope grid support film (**c** and **e**) or adsorbed to plastic slides before embedding and thin sectioning (**d**). In either case, only periplasmic epitopes were labelled, indicating that both faces of the two-dimensional crystals correspond to the periplasmic surface of the complex. Scale bars: **a**, **b**, 500 nm; **c–e**, 200 nm.

Therefore, helices 14 and 15 may correspond to SecG. The remaining 13 helices can be described in terms of a central four-helix bundle (tilted helices 6–9) with a left-handed twist, surrounded by three other groups: a layer of less tilted helices (11–13) on one side, a right-handed four-helix bundle (helices 1–4) on the opposite side, and two highly tilted helices (5 and 10) at the dimer interface (Fig. 3a).

In the centre of the SecYEG dimer there is a clear cavity 16 × 25 Å wide and 22 Å deep around the position of the noncrystallographic two-fold axis (asterisk in Fig. 1). This indentation is closed on the periplasmic membrane surface (Fig. 1), and is formed at the interface of the two monomers. The cavity is lined by the whole length of helices 5 and 10 and, on the cytoplasmic half, also by helices 6 and 9. The highly tilted helix 10 is in close contact with the same helix of the adjacent monomer on the periplasmic membrane side. These two helices tilt away from one another towards the cytoplasmic side in a V-shaped configuration (Fig. 3a, b), closing the cavity on the periplasmic side.

Our map shows that a single SecYEG monomer does not contain a channel that would be large enough for protein translocation. Consistent with this result, most studies of SecYEG and Sec61p suggest that the translocation channel is formed by an assembly of several copies of the translocase complex, although the exact stoichiometry is uncertain. A recent analysis found that translocating protein is associated with a dimer of SecYEG¹⁴, but trimers⁷ and tetramers^{5,8} have also been proposed to form the translocation conduit. All these studies suggest, however, that the channel is

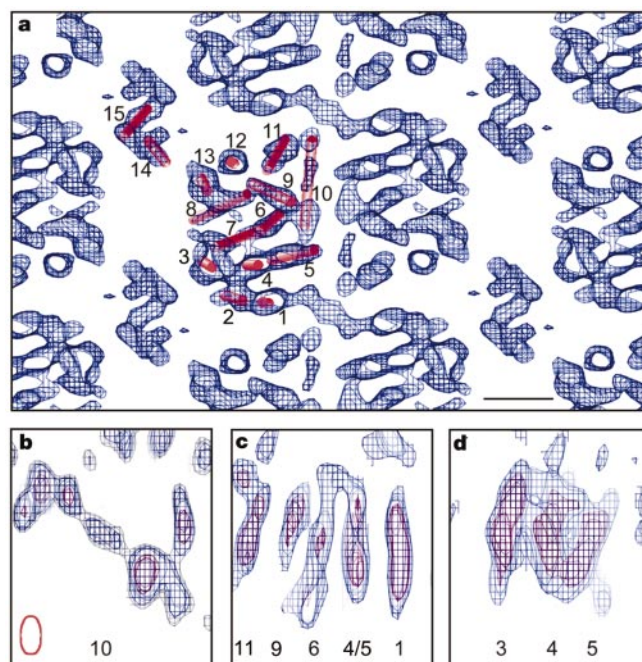


Figure 3 Dimeric structure of the SecYEG complex. **a**, Top view from the cytoplasmic side of one crystalline membrane with applied two-fold noncrystallographic symmetry. The unsymmetrized map and the electron density file (as a brk input to the program O³⁰) are available as Supplementary Information. SecYEG dimers are clearly separated from one another. Within one monomer, the 15 transmembrane helices are drawn as red cylinders. The two peripheral helices, 14 and 15, may correspond to the non-essential SecG subunit. **b–d**, Side views of helices contoured at 1.0 (grey; in **b** only), 1.35 (blue) and 2.7 (pink) s.d. In **c**, tilted helix 9 extends outside the slab. All helix densities are continuous at one s.d. At increasing contour levels, the densities of the most highly tilted helices, 10 and 7, appear discontinuous, as expected owing to the missing cone of data. The anisotropy of the resolution is illustrated by the elongation of the point spread function along the direction perpendicular to the membrane plane²⁸ (red inset in **b**).

Table 1 **Crystallographic data summary**

Plane group symmetry	$p12_1-b$
Unit cell dimensions	$a = 104 \text{ \AA}, b = 57 \text{ \AA}, \gamma = 90^\circ$
Thickness	$\sim 120 \text{ \AA}$
Number of lattices*	63
Total no. of observations†	6,213
No. of unique structure factors‡	2,175
Overall weighted phase error†§	20.3°

*Distribution of tilt angles: 0–10°, 17 images; 10–20°, 8 images; 20–30°, 6 images; 30–40°, 26 images; 40–55°, 6 images.

†Including reflections with signal-to-noise ratios greater than one (IQ 1–8; ref. 25).

‡Before imposing the noncrystallographic symmetry.

§90° is random.

located at the interface between monomers. Our map of the SecYEG dimer shows a deep cavity at the dimer interface that is closed at the periplasmic side of the membrane. This may be the closed translocation channel. The diameter of the cavity is similar to that of previously observed translocation pores, ranging from 9 to 25 Å (refs 5, 7, 8). Furthermore, the most recent map of a ribosome-associated Sec61 channel also displays a central indentation rather than a continuous channel, and is closed on the side of the membrane that would correspond to the periplasmic surface⁷. Small conformational changes would be sufficient to convert the observed cavity into a continuous channel: the helices 10 of adjacent monomers could move away from one another to facilitate protein translocation. Helix 10 could correspond to the conserved third transmembrane domain of SecE, which has been shown to be a contact site between two monomers¹⁵ and is thought to have a dynamic role in protein translocation^{15–18}. However, we cannot exclude the possibility that the active channel is formed upon incorporation of additional SecYEG monomers into the oligomeric complex⁸. Irrespective of the mechanism of channel formation, the dimer we find in the membrane crystals could represent a state of the SecYEG complex that can initiate protein translocation. Therefore, it would contain the binding sites for SecA and for the signal sequence of a translocation substrate. SecA is an ATPase that delivers the protein substrate to the channel and is thought to move in and out of a SecYEG structure while being shielded from phospholipid^{19–21}. This interaction results in a progressive movement of the polypeptide through the pore^{19,20}. SecA could be bound by the extensive cytosolic loops of the SecYEG complex near the dimer interface. The cavity is too small to enclose the entire SecA dimer, which measures about 150 × 80 × 80 Å (ref. 22), suggesting that only a part of SecA inserts into the channel, or that additional SecYEG molecules combine to engulf SecA. The cavity we observe would provide an easy route for the signal sequence to proceed from SecA to then intercalate between the SecY transmembrane helices¹⁷ and thus to initiate protein translocation. □

Methods

Antibody labelling

On-grid and pre-embedding immunogold labelling were performed as described²³. Antibodies were raised against synthetic peptides corresponding to the last 15 and 16 amino acids of SecY and SecG, respectively, plus additional cysteines at the amino terminus. The sequences chosen, according to topology predictions, correspond to both periplasmic (SecG) and cytoplasmic (SecY) regions of the complex. The binding sites of the antibodies with respect to the plasma membrane were confirmed by freeze-fracture replica labelling of whole cells²⁴ (Fig. 2a, b).

Data collection, processing and analysis

Two-dimensional crystals of SecYEG were obtained and imaged by electron cryomicroscopy as described⁹. Images recorded at 0, 20, 30 or 45° nominal tilt were evaluated by optical diffraction. Well-ordered areas of 4,000 × 4,000 pixels were digitized on a Zeiss SCAI scanner with a 7-μm step. Images were corrected for distortion of the crystal lattice, the effects of the contrast transfer function (CTF) and astigmatism using the MRC image-processing programs^{25,26}. Initial estimates of tilt angles were calculated from lattice parameters, and amplitudes and phases of structure factors were merged in plane group $p12_1$. The absolute hand of the map was checked with an image of tilted purple membrane, which was compared to the bacteriorhodopsin 3D data set²⁷. The phase origin, tilt geometry and CTF of each image were refined. About 70 of ~2,000 images yielded 8–

10 Å data, and 63 lattices were merged. The vertical resolution was estimated at ~16 Å by the point-spread function²⁸. Image amplitudes were scaled with an average temperature factor of $-600 \text{ \AA}^2$ to compensate for resolution-dependent drop-off. The 3D density map was calculated with CCP4 programs²⁹, and the map displayed with the program O³⁰. The two-fold noncrystallographic symmetry was determined by the CCP4 program FINDNCS and applied using the program MAPROT.

Received 11 March; accepted 18 April 2002; doi:10.1038/nature00827.

- Matlack, K., Mothes, W. & Rapoport, T. Protein translocation: tunnel vision. *Cell* **92**, 381–390 (1998).
- Brundage, L., Hendrick, J. P., Schiebel, E., Driessen, A. J. & Wickner, W. The purified *E. coli* integral membrane protein SecY/E is sufficient for reconstitution of SecA-dependent precursor protein translocation. *Cell* **62**, 649–657 (1990).
- Akimaru, J., Matsuyama, S. I., Tokuda, H. & Mizushima, S. Reconstitution of a protein translocation system containing purified SecY, SecE, and SecA from *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **88**, 6545–6549 (1991).
- Gorlich, D., Prehn, S., Hartmann, E., Kalies, K. & Rapoport, T. A. A mammalian homolog of SEC61p and SECYp is associated with ribosomes and nascent polypeptides during translocation. *Cell* **71**, 489–503 (1992).
- Hanein, D. et al. Oligomeric ring of the Sec61p complex induced by ligands required for protein translocation. *Cell* **87**, 721–732 (1996).
- Ménétret, J.-F. et al. The structure of ribosome–channel complexes engaged in protein translocation. *Mol. Cell* **6**, 1219–1232 (2000).
- Beckmann, R. et al. Architecture of the protein-conducting channel associated with the translocating 80S ribosome. *Cell* **107**, 361–372 (2001).
- Manting, E., van der Does, C., Remigy, H., Engel, A. & Driessen, A. J. M. SecYEG assembles into a tetramer to form the active protein translocation channel. *EMBO J.* **19**, 852–861 (2000).
- Collinson, I. et al. Projection structure and oligomeric properties of a bacterial core protein translocase. *EMBO J.* **20**, 2462–2471 (2001).
- Nishiyama, K., Suzuki, T. & Tokuda, H. Inversion of the membrane topology of SecG coupled with SecA-dependent preprotein translocation. *Cell* **85**, 71–81 (1996).
- Akiyama, Y. & Ito, K. Topology analysis of the SecY protein, an integral membrane protein involved in protein export in *Escherichia coli*. *EMBO J.* **6**, 3465–3470 (1987).
- Duong, F. & Wickner, W. Distinct catalytic roles of the SecY, SecE, and SecEYfajC subunits of preprotein translocase holoenzyme. *EMBO J.* **16**, 2756–2768 (1997).
- Nishiyama, K., Hanada, M. & Tokuda, H. Disruption of the gene encoding p12 (SecG) reveals the direct involvement and important function of SecG in the protein translocation of *Escherichia coli* at low temperature. *EMBO J.* **13**, 3272–3277 (1994).
- Bessonneau, P., Besson, V., Collinson, I. & Duong, F. The SecYEG preprotein translocation channel is a conformationally dynamic and dimeric structure. *EMBO J.* **21**, 995–1003 (2002).
- Kaufmann, A., Manting, E. H., Veenendaal, A. K., Driessen, A. J. & van der Does, C. Cysteine-directed cross-linking demonstrates that helix 3 of SecE is close to helix 2 of SecY and helix 3 of a neighbouring SecE. *Biochemistry* **38**, 9115–9125 (1999).
- Flower, A. M., Osborne, R. S. & Silhavy, T. J. The allele-specific synthetic lethality of prlA-prlG double mutants predicts interactive domains of SecY and SecE. *EMBO J.* **14**, 884–893 (1995).
- Plath, K., Mothes, W., Wilkinson, B. M., Stirling, C. J. & Rapoport, T. A. Signal sequence recognition in posttranslational protein transport across the yeast ER membrane. *Cell* **94**, 795–807 (1998).
- Veenendaal, A. K., van der Does, C. & Driessen, A. J. Mapping the sites of interaction between SecY and SecE by cysteine scanning mutagenesis. *J. Biol. Chem.* **276**, 32559–32566 (2001).
- Economou, A. & Wickner, W. SecA promotes preprotein translocation by undergoing ATP-driven cycles of membrane insertion and deinsertion. *Cell* **78**, 835–843 (1994).
- van der Wolk, J. P., de Wit, J. G. & Driessen, A. J. The catalytic cycle of the *Escherichia coli* SecA ATPase comprises two distinct preprotein translocation events. *EMBO J.* **16**, 7297–7304 (1997).
- Eichler, J., Brunner, J. & Wickner, W. The protease-protected 30 kDa domain of SecA is largely inaccessible to the membrane lipid phase. *EMBO J.* **16**, 2188–2196 (1997).
- Shilton, B. et al. *Escherichia coli* SecA shape and dimensions. *FEBS Lett.* **436**, 277–282 (1998).
- Kleymann, G., Ostermeier, C., Heitmann, K., Haase, W. & Michel, H. Use of antibody fragments (Fv) in immunocytochemistry. *J. Histochem. Cytochem.* **43**, 607–614 (1995).
- Fujimoto, K. SDS-digested freeze-fracture replica labeling electron microscopy to study the two-dimensional distribution of integral membrane proteins and phospholipids in biomembranes: practical procedure, interpretation and application. *Histochem. Cell Biol.* **107**, 87–96 (1997).
- Henderson, R., Baldwin, J. M., Downing, K. H. & Zemlin, F. Structure of purple membrane from *Halobacterium halobium*. Recording, measurement and evaluation of electron micrographs at 3.5 Å resolution. *Ultramicroscopy* **19**, 147–178 (1986).
- Crowther, R. A., Henderson, R. & Smith, J. M. MRC image processing programs. *J. Struct. Biol.* **116**, 9–16 (1996).
- Grigorieff, N., Ceska, T. A., Downing, K. H., Baldwin, J. M. & Henderson, R. Electron-crystallographic refinement of the structure of bacteriorhodopsin. *J. Mol. Biol.* **259**, 393–421 (1996).
- Unger, V. M. Assessment of electron crystallographic data obtained from two-dimensional crystals of biological specimens. *Acta Crystallogr. D* **56**, 1259–1269 (2000).
- Collaborative Computational Project No. 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
- Jones, T. A., Zou, J. Y., Cowans, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps. *Acta Crystallogr.* **47**, 110–119 (1991).

Supplementary Information accompanies the paper on *Nature's* website (<http://www.nature.com/nature>).

Acknowledgements

We thank E. Or for the production of the polyclonal antibodies, F. Joos for immunostaining of thin sections, and F. Duong for critically reading the manuscript. C.B. is grateful to D. Mills for help with the JEOL3000SF electron microscope, V. Unger for advice on image processing, J. Vonck for discussions on noncrystallographic symmetry, and D. Picot for discussions and advice on crystallographic programs. I.C. was a fellow of

the Human Frontiers Science Program at Harvard University. C.B. acknowledges support from Jean-Luc Popot and the CNRS-UMR7099, where she carried out the final part of the analysis.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.B. (e-mail: Cecile.Breyton@ibpc.fr) or I.C. (e-mail: Ian.Collinson@mpibp-frankfurt.mpg.de).

corrigendum

Extensive and divergent circadian gene expression in liver and heart

Kai-Florian Storch, Ovidiu Lipan, Igor Leykin, N. Viswanathan, Fred C. Davis, Wing H. Wong & Charles J. Weitz

Nature **417**, 78–83 (2002).

Two errors of gene annotation have come to our attention. Although we refer to *Zfp36* in the text, the gene identified in our data sets was *Zfp36l-1* (*Zfp36-like 1*; NCBI RefSeq accession number NM_007564). Current evidence suggests that *Zfp36l-1* protein and its close relatives are RNA-binding factors rather than transcription factors^{1–3}. Rather than *Thra* (thyroid hormone receptor- α), which we identified on the basis of an incorrect Unigene cluster assignment in NCBI, the correct assignment is the nuclear orphan receptor *Rev-erb- β* (GenBank accession number U09504). These corrections do not affect our conclusions in any significant way.

1. Carballo, E., Lai, W. S. & Blackshear, P. J. Feedback inhibition of macrophage tumor necrosis factor- α production by tristetraprolin. *Science* **281**, 1001–1005 (1998).
2. Lai, W. S. *et al.* Evidence that tristetraprolin binds to AU-rich elements and promotes the deadenylation and destabilization of tumor necrosis factor α mRNA. *Mol. Cell Biol.* **19**, 4311–4323 (1999).
3. Lai, W. S., Carballo, E., Thorn, J. M., Kennington, E. A. & Blackshear, P. J. Interactions of CCHC zinc finger proteins with mRNA. Binding of tristetraprolin-related zinc finger proteins to Au-rich elements and destabilization of mRNA. *J. Biol. Chem.* **275**, 17827–17837 (2000).

addendum

Nitrogen loss from unpolluted South American forests mainly via dissolved organic compounds

Steven S. Perakis & Lars O. Hedin

Nature **415**, 416–419 (2002).

In this Letter we reported that hydrologic export of dissolved organic nitrogen (DON) dominates over nitrate in unpolluted old-growth forests across southern Chile and Argentina, but that the reverse pattern occurs in old-growth forests exposed to chronically high rates of nitrogen deposition in eastern North America. As a note of clarification, however, we feel it is useful to point out that, depending on conditions, second-growth forests (that is, those that have been previously logged) can display various patterns of nitrogen loss, including dominance of DON over nitrate^{1,2} in cases where forest regrowth and detritus accumulation exert strong and well-known demands on internal nitrate supply^{3–6}. But these forests were not included in our analysis, given that their nitrogen cycles are influenced by historically complex interactions between nitrogen deposition and land use and so do not represent an appropriate comparison to the old-growth forests investigated in South America.

1. Goodale, C. L., Aber, J. D. & McDowell, W. H. The long-term effects of disturbance on organic and inorganic nitrogen export in the White Mountains, New Hampshire. *Ecosystems* **3**, 433–450 (2000).
2. Campbell, J. L. *et al.* Dissolved organic nitrogen budgets for upland, forested ecosystems in New England. *Biogeochemistry* **49**, 123–142 (2000).
3. Vitousek, P. M. & Reiners, W. A. Ecosystem succession and nutrient retention: a hypothesis. *BioScience* **25**, 376–381 (1975).
4. Hedin, L. O., Armesto, J. J. & Johnson, A. H. Patterns of nutrient loss from unpolluted, old-growth temperate forests: evaluation of biogeochemical theory. *Ecology* **76**, 493–509 (1995).
5. Likens, G. E. & Bormann, F. H. *Biogeochemistry of a Forested Ecosystem* 2nd edn (Springer, New York, 1995).
6. Aber, J. D. *et al.* Nitrogen saturation in temperate forest ecosystems: Hypotheses revisited. *BioScience* **48**, 921–934 (1998).

Recent introductions

Wipe that DNA right out of your plastic and glassware containers.

Multiskan Spectrum

Thermo Labsystems www.thermo.com

A new UV/VIS microplate and cuvette spectrophotometer

The Multiskan Spectrum combines both standard UV/visible spectrophotometer cuvette reading capabilities and advanced microtitre plate reading into a convenient, compact instrument. The dual-beam cuvette reader offers both sample and reference beams, for enhanced, long-term stability. Multiskan Spectrum can make quick absorbance measurements in a cuvette between the cycles of a microtitre plate enzyme kinetics experiment. The cuvette port's short measurement cycle optimizes enzyme kinetics conditions before transferring kinetic protocols to microtitre plate format.

Montage DNA Gel Extraction Kit

Millipore www.millipore.com

A ten-minute spin yields fully functional DNA

The latest addition to the Montage range of genomics sample prep kits is Montage DNA Gel Extraction Kit for the recovery of fully functional DNA from agarose gel slices in a single 10-minute spin. DNA fragments that are 100–10,000 base pairs in size may be efficiently recovered ready for sequencing, cloning or labelling with no additional purification. The Montage DNA Gel Extraction kit is available as a 50-pack containing 50 Montage DNA Gel Extraction (formerly Ultrafree-DA) centrifugal devices and, for added convenience, a bottle of 500 ml of 50× modified TAE buffer for gel casting and running. Each centrifugal device is ready to use, consisting of three pre-assembled components: a gel nebulizer, a sample filter cup and a collection microcentrifuge tube. The specially formulated modified TAE buffer contains one-tenth the standard EDTA concentration to minimize interference with downstream applications.



Montage: 50 ways to yield your DNA.

MiniBeadbeater-96

Biospec Products www.biospec.com

High-capacity cell disrupter

Designed for high-throughput analytical screening, the MiniBeadbeater-96 will disrupt up to 128 samples of spores, microorganisms, plant and animal tissue or soil samples at a time. Using the same homogenization process as its predecessors, MiniBeadbeater-1 and -8, one or two deep-well microtitre plates containing sample and a large number of tiny ceramic beads are violently shaken for 2–3 minutes to achieve over 90% cell disruption. The plates and beads are disposable, making them suitable for PAGE, PCR and Probe applications where cross-contamination between samples cannot be tolerated. Compatible with phenol, Gu-SCN and commercial kit solutions, RNA/DNA is quickly recovered at high yield by doing cell disruption and nucleic acid extraction simultaneously.



A bit shaky: MiniBeadbeater-96.

TRobot

Biometra www.biometra.com

Let the machines take the strain

The Biometra TRobot is designed specifically for integration in robotic systems. The thermocycler module is separated from the controller unit for a small footprint. The motorized heated lid ensures automated workflow and reliable sealing of plates. TRobot features a motorized plate lifter. Once lifted the plate can be easily removed by a robotic arm. All features of the TRobot are controlled over its serial port. Integration in existing software environments (of individual robotic stations) is very easy. In addition to controlling standard features, the TRobot can provide extensive status messages. A program for operating up to 16 TRobots from a computer is also available. The TRobot comes with 96- or 384- well block.

BD TALON HT

BD Biosciences Clontech www.clontech.com

In His name

The BD TALON HT 96-Well Plate is intended for studies needing rapid or high-throughput purification of histidine-tagged proteins. Each well plate is pre-loaded with 200 mg of BD TALON Superflow Resin in each well of a 96-well filtration plate. That provides enough resin to bind up to 1 mg of histidine-tagged protein per well. The plate is provided with convenient top and bottom seals for shipping and storage, as well as a collection plate that can be used for receiving the eluted His-tagged proteins and for storage of samples.

NucleasEliminator wipes

AMRESCO www.amresco-inc.com

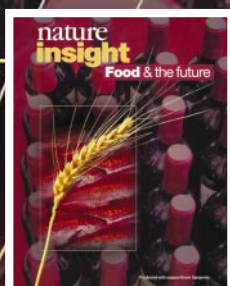
Nucleases removed at a stroke

NucleasEliminator Wipes are a handy and convenient way to remove RNase and DNase contamination. The wipes are ready to use, single-use moist towelettes packed singly in a foil pouch. Active ingredient is AMRESCO's NucleasEliminator solution. Suitable for removing nuclease from any plastic and glassware surfaces without contaminating the experiment, these wipes are recommended for researchers performing PCR and delicate electrophoresis reactions and experiments. Each box contains 25 wipes and can be stored at room temperature.

These notes are compiled in the Nature office from information provided by the manufacturers.

nature insight

Food & the future



Cover illustration
courtesy of Getty
Images.

Bovine spongiform encephalopathy, genetically modified (GM) crop plants, antibiotics and hormones in animal feed, the impact of agriculture on biodiversity, fast food and globalization, famine in Africa — food is news. One strand unites these issues, and it can be summarized in one word: 'sustainability'. The world's population continues to grow, yet resources are finite. Our mission is to squeeze more crops from the same patch of ground, while preserving that patch in a state fit to pass on to further hungry generations. The quest for sustainability is the theme of this Insight.

We have been here before. After the Second World War, the application of science created a range of high-yield crops, averting worldwide famine. Half a century on, this Green Revolution is played out and the world needs yet greater ingenuity to feed itself. Once again, science is at the sharp end. In the West, especially in crowded Europe, public debate has focused on GM crops. These offer one of many ways to feed the world in the coming decades, and it would be wrong to concentrate on GM crops to the exclusion of other issues that affect the future of food production, such as water availability, intelligent land use, waste management and climate change.

Not that we should ignore concerns about the environmental impact of GM crops (see, for example, the June 2002 issue of *Nature Biotechnology*). However, as this collection illustrates, GM crops form only one item in a picture of agriculture shaped by a wider, historical perspective. Farming has shaped the world, and ourselves, for more than ten millennia. The decisions we make now could affect humanity for millennia to come.

We are pleased to acknowledge the financial support of Syngenta in producing this Insight. As always, *Nature* carries sole responsibility for all editorial content and peer review.

Henry Gee Senior Editor

Editor, *Nature*: Philip Campbell
Insights Editor: Lesley Anson
Editorial Assistant: Simon Gibson
Production Editor: Simon Gribbin
Art Director: Majo Xeridat

Diagrams: Ann Thomson
Vicky Askew
Suzanne Coleman
Production: Jeffrey Bird
Marketing: Claire Aspinall
Sponsorship: Emma Jones

overview

- 668 Malthus foiled again and again
A. Trewavas

review articles

- 671 Agricultural sustainability and intensive production practices
D. Tilman, K. G. Cassman, P. A. Matson, R. Naylor & S. Polasky

- 678 Enhancing the crops to feed the poor

J. Huang, C. Pray & S. Rozelle

progress

- 685 Assessing the risks associated with new agricultural practices
R. S. Hails

review articles

- 689 Towards sustainability in world fisheries
D. Pauly, V. Christensen, S. Guénette, T. J. Pitcher, U. R. Sumaila, C. J. Walters, R. Watson & D. Zeller

progress

- 696 The present and future of the international wine industry
L. F. Bisson, A. L. Waterhouse, S. E. Ebeler, M. A. Walker & J. T. Lapsley

review articles

- 700 Evolution, consequences and future of plant and animal domestication
J. Diamond

Malthus foiled again and again

Antony Trewavas

Institute of Cell and Molecular Biology, University of Edinburgh, Mayfield Road, Edinburgh EH9 3JH, UK (e-mail: trewavas@ed.ac.uk)

Throughout history, increasing population has driven the need to increase agricultural efficiency, so averting successive 'malthusian' disasters. In the twentieth century, the application of scientific knowledge to agriculture yielded tremendous dividends, enabling cereal yields to increase threefold since 1950. But with the world's population projected to reach nine billion by the middle of this century, new ways must be found to increase yields while preserving natural habitats and biodiversity.

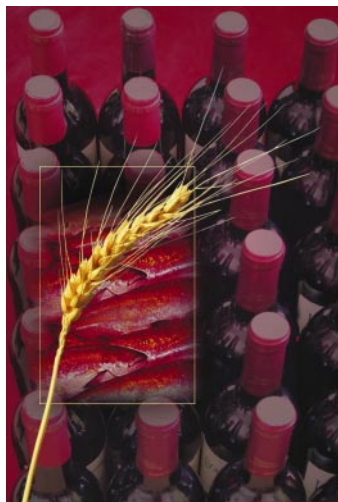
Ever since the Reverend Thomas Malthus published *An Essay on the Principle of Population* in 1798 there has been concern over the twin problems of population and food¹. Malthus, at least in his first edition, foresaw a human population always hungry and therefore malnourished. To be fair to Malthus, in later editions he did back-track on this apocalyptic view of the future of humanity, but he did receive the ultimate accolade — the incorporation of his name into the English language, with 'malthusian' now a recognized adjective.

At the time Malthus wrote and lived, famine was common in Europe. The French revolution in 1789 resulted from harvest failures and a consequent steep rise in the price of bread. The amount of land that was farmed was determined by how much wheat could be harvested during the three- to four-week harvesting period by one person with a scythe. The development of agricultural engineering led quickly to horse-drawn rotary harvesters that increased the amount of wheat that could be harvested, and farms grew as a result. Current agricultural technology enables one person to be fed from the food grown on no more than 2,000 square metres². In Malthus's time it was nearer 20,000 square metres.

An increasing population drove the need to increase agricultural efficiency even further. In the middle of the nineteenth century, Leibig and others established the mineral content of plants and thus the chemicals needed to provide for good crop growth. Although the world population at the time of the French revolution was about one billion², the twin effects of engineering and fertilizer development enabled a doubling by the middle of the nineteenth century. Sources of fertilizer at that time were either deposits of guano (seabird droppings) or potassium nitrate (saltpetre) and by the end of the century many of these reserves had been fully exploited. Concerns grew that a burgeoning world population would once again be plunged into malthusian disaster.

A century of change

Salvation came from one of the few beneficial offshoots of the First World War: the development of the Haber–Bosch process in which atmospheric nitrogen was fixed and used



to manufacture ammonia². The Haber process was a product of the explosives industry (ammonia is used to make dynamite), but the immediate consequence was that humanity no longer had to rely on finite natural resources of nitrogenous fertilizer. Even coupling the Haber process to renewable energy sources has the potential to provide unlimited supplies of nitrogenous fertilizer — global output of fixed nitrogen from the chemical industry currently exceeds that of the biological nitrogen cycle. But overuse has its disadvantages, including the pollution of waterways, lakes and seas with nitrate, the encouragement of toxic algal blooms and the deterioration of water supplies³.

The twentieth century was the era of plant breeding and genetics. The introduction of the concept of 'hybrid vigour' by Schull in 1900 stabilized US corn production⁴. For the first time the genetic base of crop production could be controlled and easily adjusted to accommodate the differing climates found in the continental United States. A century later, however, this process has led to the predominance of monocultures of limited genetic variability and increased vulnerability to the threat of disease. Malthusian crisis was averted once again with the 'Green Revolution' of the 1950s, by the introduction of dwarfing genes into rice and wheat⁴. Cereal yield over the past half century has increased threefold, allowing the human population to reach six billion.

No room for complacency

The application of scientific knowledge to agriculture has yielded extraordinary dividends. Although estimates suggest that about 800 million people are still undernourished, it is thought that this number will drop to about 600 million, largely in sub-Saharan Africa, by 2025 (refs 1, 2). But this is no time for congratulation: although it is hoped the human population will level off at about nine billion by 2050, the population is currently still expanding. Additionally, as populations get richer, meat consumption increases and, because cattle are fed largely on cereals, cereal yields will have to at least double to keep pace.

Achieving this target will face an additional constraint not seen before — lack of available farmland. From 1800 onwards, more food was simply produced by ploughing up virgin land and forest. The land area used for farming

increased about fivefold up to the middle of the twentieth century in step with population increases. The Green Revolution put a brake on this expansion, increasing yields threefold with no need for further expansion⁵. Since 1950, the proportion of the land devoted to farming has barely increased, even though the world population doubled over the same period. We currently use at least half the available good quality soil for agriculture, with the remainder under tropical forests⁶. This leads to an obvious dilemma. Unless we can pull off a second Green Revolution, increasing yield but limiting it to land currently used for farming, there will be further deterioration of natural habitats and biodiversity at a rate that could even threaten the further existence of humanity.

The lessons of history are clear. Successive lurches in population number have driven the development of new agricultural technologies designed to provide food for growing populations. This process of discovery will continue until there is an abundance of food equally enjoyed by the whole world population. We are far from achieving that at the present time, and there is therefore a constant need to examine the state of current agriculture to see where progress needs to be made. The following collection of articles on 'Food and the future' provides a snapshot of the current state of play.

Maximal yields with minimal damage

The state of agriculture across the world varies from the machine-dependent industrial farming of North American prairies to the slash-and-burn method still prevalent in parts of South America and Africa. Uniform solutions are not therefore likely, except that yield on present farmland must be increased, by whatever means. But as in previous malthusian scenarios, new technologies are emerging. Tilman *et al.* (pages 671–677) point to both the merits and demerits of modern industrial agriculture. Cereal yields continue to grow, but the environmental cost of maintaining the high standards of living to which people in the developed world have become accustomed is a cause for concern.

There have been a number of ingenious suggestions for changes in agricultural practice that would improve the environment, while at the same time increasing the efficiency of fertilizer and pesticide use, and maximizing crop yield. One example is 'precision agriculture' in which the state of the soil is monitored metre by metre, with sowing and treatment rates adjusted accordingly. Another is the 'no-till' strategy in which soil structure and biodiversity is conserved by obviating the need for ploughing. Tilman *et al.* make the additional and unusual suggestion that farmers are rewarded for environmental friendliness. An alternative is to point out to economically minded farmers that excessive use of fertilizer and pesticides, as well as being environmentally damaging, also represents a financial loss. Poor management is ultimately responsible for many agricultural problems.

One new and controversial technology is the genetic manipulation of crops, which has the potential to unlock substantial increases in yield. The use of genetically modified (GM) crops has already boosted food and fibre production, raised farm incomes and reduced pesticide usage in the countries that grow them. Huang *et al.* (pages 678–684) describe the Chinese experience of GM crops in the context of the benefits of the technology to the poorest farmers, who have the most to gain provided that GM seed is provided free or at controlled prices, much as in the Green Revolution. Thus the Chinese government has largely kept GM technology to itself, with Chinese universities and institutes developing many new GM crops.

The benefit of GM technology to the poorest farmers is palpable. To a cotton farmer working on a farm of about a hectare in area, the use of 'Br' cotton (containing a gene for an insecticide derived from the bacterium *Bacillus thuringiensis*) has raised

income by a quarter, cut costs by a third, and slashed pesticide use by three quarters. There are concerns that the cotton bollworm — the pest whose activity *Bt* cotton is designed to curb — might evolve resistance to the insecticide, but ongoing technological development of other GM lines will almost certainly ameliorate the problem if it emerges.

Antagonism in Europe to GM technology, partly because of concerns about the potential for uncontrolled spread of transgenes into weedy or invasive plants, is likely to subside once the real benefits for the consumer emerge. There is much promise in the concept of designer foods in which 'problem' substances such as gluten or common allergenic proteins are eliminated, and useful secondary products (such as the so-called 'neutra-ceuticals') are boosted.

Assessing the impact of land-use change

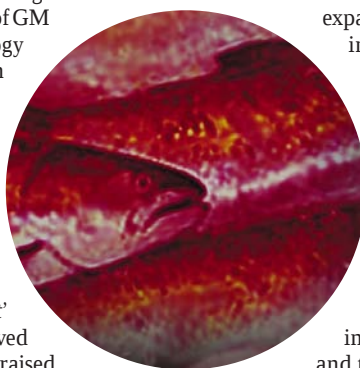
People who have a high standard of living are reluctant to lose it. But fear of change will have to be mastered, if only because the world refuses to stand still, and constant social adjustments are necessary to accommodate new or long-standing problems. Hails (pages 685–688) discusses the present attitudes to risk and risk assessment concerning crops and food and sets the scene primarily in a European context where dense human populations leave little extra land for biodiversity and leisure. Thus in the United Kingdom, farmland is a major repository of the present, limited biodiversity and its survival is highly dependent on the precise forms of land management used. The farmer is now called upon to fulfil the conflicting roles of steward of the countryside and provider of food, while his or her activities come under increasing scrutiny.

Agriculture has repeatedly met malthusian watersheds — and has overcome them. The current malthusian crisis affecting world fisheries has been caused by the application of modern harvesting to an industry that has yet to escape the ethos of the hunter-gatherer. Pauly *et al.* (pages 689–695) describe the stark situation. Mining the sea at the present rate will result in the extinction of fish species and the destruction of whole ecosystems, with unforeseen consequences. Without massive reductions of fishing fleets, these very real environmental disasters could spread onto the land. Because fish is regarded as a healthy source of protein, fish farming will have to expand substantially, but current research resources to generate the methods to farm marine fish are limited and need considerable investment. A problem is that fish-farming as currently practised is not sustainable, consuming more fish protein than it produces and leaving environmental disaster in its wake.

The fishing industry may be reluctant to change, but this is not true for all food production. The production of wine is a case in point.

As Bisson *et al.* show (pages 696–699), winemaking has expanded from a cottage-garden pursuit to a global industry in a matter of a few decades, while remaining responsive to the changing tastes of consumers. Although often thought as an indulgence of the rich, winemaking is increasingly important in the economies of many countries outside the developed world. Oenology is humankind's earliest biotechnology, and stands out as a good example of how the food industry can face the future.

How then did humans end up using agriculture? Diamond (pages 700–707) describes the intriguing history of crop and animal domestication and the forces that changed bands of hunter-gatherers



into settled agricultural communities. One important factor is the high birth rate that a settled lifestyle allows, leading to city life and the diversification of activities, while driving hunter–gatherers to the margins. But why did agriculture begin when it did — simultaneously in several different parts of the world, around 10,000 years ago? Climatic change at the end of the last Ice Age may have been the synchronizing factor, although many questions remain about this crucial episode in human history. However domestication began, the long-term benefits eventually outweighed the early disadvantages, such as the emergence of diseases that capitalized on crowded, settled populations. Only now can we really see how the creative achievements in the arts, literature, sciences and architecture could never have happened had the hunter–gatherer model predominated.

Back to the future

From the perspective of ten thousand years, however, some look back to the hunter–gatherer lifestyle with wistful nostalgia. They argue for a retreat from modern technology, so that humankind can achieve some kind of balance with nature. The deleterious environmental

consequences of some kinds of food production are very real. This is clearly evident in fisheries, and incorrectly perceived as such by some in the adoption of GM technology. But nostalgia isn't what it used to be: organic farming, sometimes touted as a panacea, is no more sustainable than the fish-farming that produces high-value smoked salmon to those consumers who can afford it. In the world at large, technological change is — as it always has been — driven by the need to squeeze ever greater yields from the same plot of land. In all such arguments, knowledge is the ultimate decider, balanced as usual by economic considerations. Whatever the outcome, the decisions we make now could have repercussions for millennia. □

doi:10.1038/nature01013

1. Dyson, T. *Population and Food—Global Trends and Prospects* (Routledge, New York, 1996).
2. Evans, L. T. *Feeding the Ten Billion—Plants and Population Growth* (Cambridge Univ. Press, Cambridge, 1998).
3. Smil, V. *Feeding the World—A Challenge for the 21st century* (MIT Press, Cambridge, MA, 2000).
4. Trewavas, A. J. The population/biodiversity paradox. Agricultural efficiency to save wilderness. *Plant Physiol.* **125**, 174–179 (2001).
5. Chrispeels, M. J. & Sadava, D. E. *Plants, Genes & Agriculture* (Jones and Bartlett Publishers, Boston, 1994).
6. Gregory, P. J. *et al.* Environmental consequences of alternative practices for intensifying crop production. *Agric. Ecosyst. Environ.* **88**, 279–290 (2002).

Agricultural sustainability and intensive production practices

David Tilman*, Kenneth G. Cassman†, Pamela A. Matson§, Rosamond Naylor|| & Stephen Polasky†

*Department of Ecology, Evolution and Behavior, and †Department of Applied Economics, University of Minnesota, St Paul, Minnesota 55108, USA (e-mail: tilman@umn.edu)

‡Department of Agronomy and Horticulture, University of Nebraska, Lincoln, Nebraska 68583, USA

§Department of Geological and Environmental Sciences, and ||Center for Environmental Science and Policy, Stanford University, Stanford, California 94305, USA

A doubling in global food demand projected for the next 50 years poses huge challenges for the sustainability both of food production and of terrestrial and aquatic ecosystems and the services they provide to society. Agriculturalists are the principal managers of global useable lands and will shape, perhaps irreversibly, the surface of the Earth in the coming decades. New incentives and policies for ensuring the sustainability of agriculture and ecosystem services will be crucial if we are to meet the demands of improving yields without compromising environmental integrity or public health.

The benefits of agriculture have been immense. Before the dawn of agriculture, the hunter-gatherer lifestyle supported about 4 million people globally¹. Modern agriculture now feeds 6,000 million people. Global cereal production has doubled in the past 40 years (Fig. 1a), mainly from the increased yields resulting from greater inputs of fertilizer, water and pesticides, new crop strains, and other technologies of the 'Green Revolution'^{2–4}. This has increased the global per capita food supply², reducing hunger, improving nutrition (and thus the ability of people to better reach their mental and physical potential) and sparing natural ecosystems from conversion to agriculture⁵.

By 2050, global population is projected to be 50% larger than at present and global grain demand is projected to double^{6–8}. This doubling will result from a projected 2.4-fold increase in per capita real income and from dietary shifts towards a higher proportion of meat (much of it grain-fed) associated with higher income. Further increases in agricultural output are essential for global political and social stability and equity. Doubling food production again, and sustaining food production at this level, are major challenges^{8–11}. Doing so in ways that do not compromise environmental integrity^{4,12,13} and public health^{14,15} is a greater challenge still. We focus here on scientific and policy challenges that must be met to sustain and increase the net societal benefits of intensive agricultural production.

Sustainability and net benefits

Agricultural practices determine the level of food production and, to a great extent, the state of the global environment. Agriculturalists are the chief managers of terrestrial 'useable' lands, which we broadly define as all land that is not desert, tundra, rock or boreal. About half of global usable land is already in pastoral or intensive agriculture⁴. In addition to causing the loss of natural ecosystems, agriculture adds globally significant and environmentally detrimental amounts of nitrogen and phosphorus to terrestrial ecosystems^{12,13}, at rates that may triple if past practices are used to achieve another doubling in food production^{4,16}. The detrimental environmental impacts

of agricultural practices are costs that are typically unmeasured and often do not influence farmer or societal choices about production methods.

Such costs raise questions about the sustainability of current practices. We define sustainable agriculture as practices that meet current and future societal needs for food and fibre, for ecosystem services, and for healthy lives, and that do so by maximizing the net benefit to society when all costs and benefits of the practices are considered. If society is to maximize the net benefits of agriculture, there must be a fuller accounting of both the costs and the benefits of alternative agricultural practices, and such an accounting must become the basis of policy, ethics and action. Additionally, the development of sustainable agriculture must accompany advances in the sustainability of energy use, manufacturing, transportation and other economic sectors that also have significant environmental impacts.

Ecosystem services

Society receives many benefits, called ecosystem services¹⁷, from natural and managed ecosystems. Ecosystems provide food, fibre, fuel and materials for shelter; additionally they provide a range of benefits that are difficult to quantify and have rarely been priced^{18,19}. Intact forests can minimize flooding by slowing snowmelt and water discharge, moderate regional climate, and remove and store atmospheric carbon dioxide, a greenhouse gas. Forest and grassland ecosystems can create or regenerate fertile soils, degrade plant litter and animal wastes, and purify water, and this regenerative process is essential for subsistence slash-and-burn farming systems²⁰. The recharge of streams and aquifers by intact ecosystems provides potable water for little more expense than the cost of its extraction.

Agricultural practices can reduce the ability of ecosystems to provide goods and services. For example, high applications of fertilizers and pesticides (Fig. 1b, c) can increase nutrients and toxins in groundwater and surface waters, incurring health and water purification costs, and decreasing fishery and recreational values. Agricultural practices that degrade soil quality contribute to eutrophication of aquatic habitats and may necessitate the expense of increased fertilization, irrigation and energy to maintain

productivity on degraded soils⁶. Practices that change species composition or reduce biodiversity in non-agricultural systems may also diminish goods and services, because the ability of ecosystems to provide some services depends both on the number and type of species in an ecosystem^{21–23}.

Global land management

The supply of agricultural products and ecosystem services are both essential to human existence and quality of life. However, recent agricultural practices that have greatly increased global food supply have had inadvertent, detrimental impacts on the environment and on ecosystem services, highlighting the need for more sustainable agricultural methods.

In the following sections, we elaborate on the benefits and costs of intensive agricultural practices that might be used to double global grain production, and suggest alternatives that might increase the net returns to society. For brevity, we do not consider the large diversity of other crops that are also critically important sources of food, incomes and agroecological stability, especially in less developed countries^{24,25}. Although our examples focus on cereal crops, livestock and practices of more developed countries, the approach to sustainability that we articulate should be relevant to all countries. However, the costs and benefits of various agricultural practices must be based on local values and local constraints, causing sustainable practices to be region and culture specific.

Although we can offer only a qualitative treatment of costs and benefits here, we believe that accurate quantification of the benefits of ecosystem services and the impact of agricultural practices on them is essential for identifying options that will lead to a more sustainable agriculture^{18,19}. Fundamental shifts in institutions, policies and incentives will be required in the search for, and broad adoption of, sustainable agricultural practices, and this search must be an on-going and adaptive process.

Food production and environmental costs

There is a general consensus that agriculture has the capability to meet the food needs of 8–10 billion people while substantially decreasing the proportion of the population who go hungry^{5,26–28}, but there is little consensus on how this can be achieved by sustainable means. Sustainability implies both high yields that can be maintained, even in the face of major shocks²⁹, and agricultural practices that have acceptable environmental impacts. The main environmental impacts of agriculture come from the conversion of natural ecosystems to agriculture, from agricultural nutrients that pollute aquatic and terrestrial habitats and groundwater, and from pesticides, especially bioaccumulating or persistent organic agricultural pollutants. Agricultural nutrients enter other ecosystems through leaching, volatilization and the waste streams of livestock and humans. Pesticides can also harm human health, as can pathogens, including antibiotic-resistant pathogens associated with certain animal production practices.

How can such costs be minimized at the same time that food production is increased? In one sense the answer is simple: crop and livestock production must increase without an increase in the negative environmental impacts associated with agriculture, which means large increases in the efficiency of nitrogen, phosphorus and water use, and integrated pest management that minimizes the need for toxic pesticides. In reality, achieving such a scenario represents one of the greatest scientific challenges facing humankind because of the trade-offs among competing economic and environmental goals, and inadequate knowledge of the key biological, biogeochemical and ecological processes.

Increasing yields

Raising yields on existing farmland is essential for 'saving land for nature', but the prospects for yield increases comparable to those of the past 40 years (Fig. 2a) are unclear^{3,10,30}. Most of the best quality

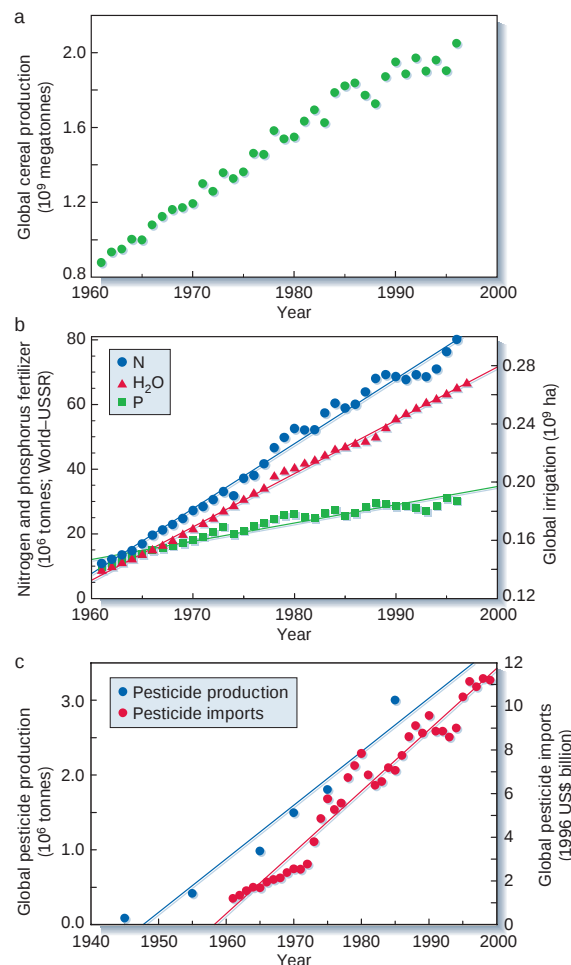


Figure 1 Agricultural trends over the past 40 years. **a**, Total global cereal production²; **b**, total global use of nitrogen and phosphorus fertilizer (except former USSR not included) and area of global irrigated land; and **c**, total global pesticide production³ and global pesticide imports (summed across all countries)². Parts **b** and **c** modified from ref. 4.

farmland is already used for agriculture, which means that further area expansion would occur on marginal land that is unlikely to sustain high yields and is vulnerable to degradation^{6,31}. Water, already limiting in many areas, may be diverted to uses that compete with irrigation. In some of the major grain production areas of east and southeast Asia, the rate of increase in rice yields is declining as actual crop yields approach a ceiling for maximal yield potential³². Finally, continuous cereal production systems, including systems with two or three crops per year, may become progressively susceptible to diseases and insect pests because of insufficient diversity in the crop rotation.

Yields have been stagnant for 15–20 years in those rice producing regions of Japan, Korea and China where farmers were early adopters of green-revolution technologies; average yields are currently about 80% of the climate-adjusted genetic yield potential ceiling³³. Lack of a larger exploitable 'yield gap' highlights the need for efforts to steadily increase the yield potential ceiling. The large yield gap for rice in many parts of south and southeast Asia, and for maize in developed and developing countries, indicates that these regions could have significant yield increases with use of appropriate technologies. Although breeders have been successful in increasing the yield

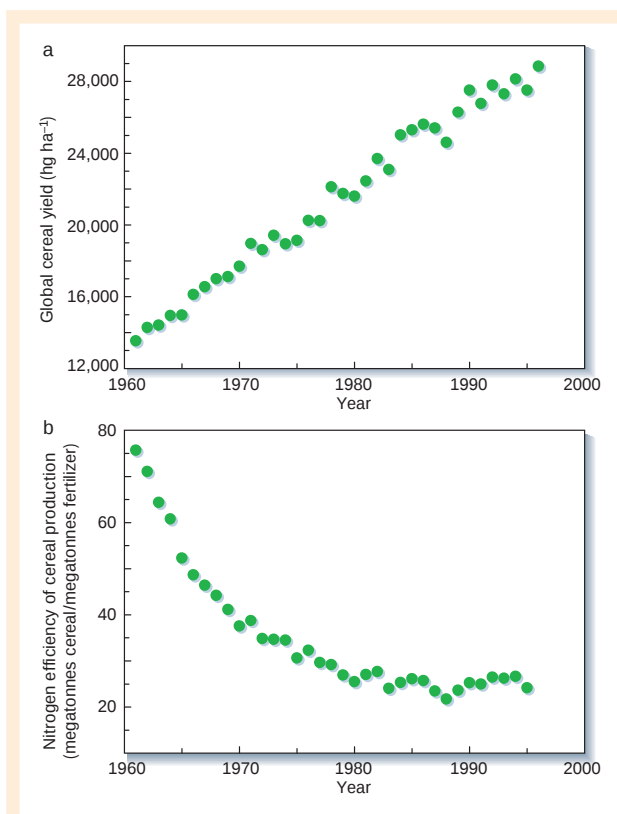


Figure 2 Diminishing returns of fertilizer application imply that further applications may not be as effective at increasing yields. **a**, Trends in average global cereal yields; **b**, trends in the nitrogen-fertilization efficiency of crop production (annual global cereal production divided by annual global application of nitrogen fertilizer)².

potential of wheat³⁴, that of inbred rice has not increased since the release of IR8 in 1966 (ref. 35), and that of maize has barely increased in 35 years³⁶. Stagnant yield potential is one of the chief impediments to sustainable agriculture and concerted efforts are needed to increase the yield potential of the major staple food crops.

Increasing nutrient-use efficiency

Intensive high-yield agriculture is dependent on addition of fertilizers, especially industrially produced NH_4 and NO_3 . In some regions of the world, crop production is still constrained by too little application of fertilizers³⁷. Without the use of synthetic fertilizers, world food production could not have increased at the rate it did and more natural ecosystems would have been converted to agriculture. Between 1960 and 1995, global use of nitrogen fertilizer increased sevenfold, and phosphorus use increased 3.5-fold (Fig. 1b); both are expected to increase another threefold by 2050 unless there is a substantial increase in fertilizer efficiency^{4,16}. Fertilizer use and legume crops have almost doubled total annual nitrogen inputs to global terrestrial ecosystems^{38,39}. Similarly, phosphorus fertilizers have contributed to a doubling of annual terrestrial phosphorus mobilization globally¹³.

Further increases in nitrogen and phosphorus application are unlikely to be as effective at increasing yields (Fig. 2a) because of diminishing returns (Fig. 2b). All else being equal, the highest efficiency of nitrogen fertilizer is achieved with the first increments of added nitrogen; efficiency declines at higher levels of addition. Today, only 30–50% of applied nitrogen fertilizer^{40,41} and ~45% of phosphorus fertilizer⁴² is taken up by crops. A significant amount of the applied nitrogen and a smaller portion of the applied phosphorus is lost from agricultural fields. This nitrogen contributes to riverine

input into the North Atlantic that is 2- to 20-fold larger than in pre-industrial times⁴³. Such non-point nutrient losses harm off-site ecosystems, water quality and aquatic ecosystems, and contribute to changes in atmospheric composition^{4,12,13,44}. Nitrogen loading to estuaries and coastal waters and phosphorus loading to lakes, rivers and streams are responsible for over-enrichment, eutrophication and low-oxygen conditions that endanger fisheries^{18,45}.

Nitrogen fertilization can increase emission of gases that have critical roles in tropospheric and stratospheric chemistry and air pollution^{46,47}. Nitrogen oxides (NO_x), emitted from agricultural soils and through combustion⁴⁸, increase tropospheric ozone, a component of smog that impacts human health, agricultural crops and natural ecosystems. As much as 35% of cereal crops worldwide are exposed to damaging levels of ozone⁴⁹. NO_x from agroecosystems can be transported atmospherically over long distances and deposited in terrestrial and aquatic ecosystems. This inadvertent fertilization can cause eutrophication, loss of diversity, dominance by weedy species and increased nitrate leaching or NO_x fluxes⁵⁰. Finally, nitrogen inputs to agricultural systems contribute to emissions of the greenhouse gas nitrous oxide. Rice paddy agriculture and livestock production are the most important anthropogenic sources of the greenhouse gas methane⁵¹.

Solutions to these problems will require significant increases in nutrient-use efficiency, that is, in cereal production per unit of added nitrogen, phosphorus and water. There are a variety of practices and improvements that could each contribute to increased efficiency. For example, nitrogen-fertilizer efficiency of maize in the United States has increased by 36% in the past 21 years as a result of large investments in public sector research and extension education, and investments by farmers in soil testing and improved timing of fertilizer application^{40,52}. The development and preferential planting of crops and crop strains that have higher nutrient-use efficiency are clearly essential. Cover crops or reduced tillage can reduce leaching, volatilization and erosional losses of nutrients and increase nutrient-use efficiency. Closing the nitrogen and phosphorus cycles, such as by appropriately applying livestock and human wastes, increases cereal production per unit of synthetic fertilizer applied.

Reliance on organic nutrient sources is a central feature of organic agriculture⁵³, but it is unclear whether the 'slow release' of nutrients from organic compost or green manures can be adequately controlled to match crop demand with nutrient supply to increase nitrogen-use efficiency in intensive cereal production systems, thereby decreasing losses to leaching and volatilization. More research on improving efficiency and minimizing losses from both inorganic and organic nutrient sources is needed to determine costs, benefits and optimal practices.

Nutrient-use efficiency is increased by better matching temporal and spatial nutrient supply with plant demand. Applying fertilizers during periods of greatest crop demand, at or near the plant roots, and in smaller and more frequent applications all have the potential to reduce losses while maintaining or improving yields and quality^{44,54–56}. Such 'precision agriculture' has typically been used in large-scale intensive farming, but is possible at any scale and under any conditions given the use of appropriate diagnostic tools⁶. Strategies that synchronize nutrient release from organic sources with plant demand are also needed^{57,58}.

Multiple cropping systems using crop rotations or intercropping (two or more crops grown simultaneously) may improve pest control and increase nutrient- and water-use efficiency. Agroforestry, in which trees are included in a cropping system, may improve nutrient availability and efficiency of use and may reduce erosion, provide firewood and store carbon.

Landscape-scale management holds significant potential for reducing off-site consequences of agriculture. Individual farms, watersheds and regional planning can take advantage of services provided by adjacent natural, semi-natural or restored ecosystems. Trees and shrubs planted in buffer strips surrounding cultivated fields

decrease soil erosion and can take up nutrients that otherwise would enter surface or ground waters. Buffer zones along streams, rivers and lakeshores can decrease nutrient and silt loading from cultivated fields or pastures. Crop pollination can be provided by insects and other animals living in nearby habitats or buffer strips, whereas other organisms from these habitats, such as parasitoids, can provide effective control of many agricultural pests. Buffer strips can also be managed to reduce inputs of weeds and other agricultural pests. The procurement of such ecosystem services will require landscape-level management.

Increasing water-use efficiency

Forty per cent of crop production comes from the 16% of agricultural land that is irrigated^{59,60}. Irrigated lands (Fig. 1b) account for a substantial portion of increased yields obtained during the Green Revolution. Unless water-use efficiency is increased, greater agricultural production will require increased irrigation. However, the global rate of increase in irrigated area is declining, per capita irrigated area has declined by 5% since 1978, and new dam construction may allow only a 10% increase in water for irrigation over the next 30 years^{60,61}. Moreover, water is regionally scarce. Many countries in a band from China through India and Pakistan, and the Middle East to North Africa either currently or will soon fail to have adequate water to maintain per capita food production from irrigated land⁶². Roughly 20% of the irrigated area of the United States is supplied by groundwater pumped in excess of recharge, and overpumping is also a serious concern in China, India and Bangladesh⁶³. Urban water use, restoration of streams for recreational, freshwater fisheries, and protection of natural ecosystems are all providing competition for water resources previously dedicated to agriculture. Finally, irrigation return-flows typically carry more salt, nutrients, minerals and pesticides into surface and ground waters than in source water, impacting downstream agricultural, natural systems and drinking water.

Technologies such as drip and pivot irrigation can improve water-use efficiency and decrease salinization while maintaining or increasing yields. They have been used in industrialized nations on high-value horticultural crops, but their expanded use currently is not economically viable for staple food crops. In developing countries, 15 million hectares have experienced reduced yields owing to salt accumulation and waterlogging⁶⁴. The water-holding capacity of soil can be increased by adding manure or reducing tillage and by other approaches that maintain or increase soil organic matter. Cultivation of crops with high water-use efficiency, and the development — through the use of biotechnology or conventional breeding — of crops with greater drought tolerance can also contribute to yield increases in water-limited production environments^{65,66}. Investment in such water-efficient technologies, however, is best facilitated when water is valued and priced appropriately.

Maintaining and restoring soil fertility

Fertile soils with good physical properties to support root growth are essential for sustainable agriculture, but, since 1945, approximately 17% of vegetated land has undergone human-induced soil degradation and loss of productivity, often from poor fertilizer and water management, soil erosion and shortened fallow periods⁶⁷. Continuous cropping and inadequate replacement of nutrients removed in harvested materials or lost through erosion, leaching or gaseous emissions deplete fertility and cause soil organic matter levels to decline, often to half or less of original levels⁴⁴. Soil tillage speeds decomposition of soil organic matter and the release of mineral nutrients. Erosion can be severe on steep slopes where windbreaks have been cleared, vegetative cover is absent during the rainy season, and where heavy machinery is involved in land preparation⁶⁴. The effects of land degradation on productivity can sometimes be compensated for by increased fertilization, irrigation, and disease control, which increase production costs⁶⁴. Crop rotation, reduced

tillage, cover crops, fallow periods, manuring and balanced fertilizer application can help maintain and restore soil fertility.

Disease and pest control

Improvements in the control of weedy competitors of crops, crop diseases and pathogens, and herbivores could significantly increase yields. Three cereals — wheat, rice and corn — provide 60% of human food. These crops, derived from once-rare weedy species, have become the three most abundant plants on Earth. A central conclusion of epidemiology is that both the number of diseases and the disease incidence should increase proportional to host abundance, and this disconcerting possibility illustrates the potential instability of a global strategy of food production in which just three crops account for so high a proportion of production. The relative scarcity of outbreaks of diseases on these crops is a testament to plant breeding and cultivation practices. For all three cereals, breeders have been successful at improving resistances to abiotic stresses, pathogens and diseases, and at deploying these defences in space and time so as to maintain yield stability despite low crop diversity in continuous cereal systems. However, it is unclear if such conventional breeding approaches can work indefinitely. Both integrated pest management and biotechnology that identifies durable resistance through multiple gene sources should play increasingly important roles^{66,68}.

Nonetheless, the evolutionary interactions among crops and their pathogens mean that any improvement in crop resistance to a pathogen is likely to be transitory. Each defence sows the evolutionary seeds of its own demise⁶⁹. Maize hybrids in the United States now have a useful lifetime of about 4 years, half of what it was 30 years ago. Similarly, agrochemicals, such as herbicides, insecticides, fungicides and antibiotics, are also major selective agents. Within about one or two decades of the introduction of each of seven major herbicides, herbicide-resistant weeds were observed⁶⁹. Insects often evolve resistance to insecticides within a decade. Resistant strains of bacterial pathogens appear within 1–3 years of the release of many antibiotics. But the need to breed for new disease resistance and to discover new pesticides can be reduced by crop rotation and the use of spatial or temporal crop diversity. Recently, an important and costly pathogen of rice was controlled in a large region of China by planting alternating rows of two rice varieties⁷⁰. This tactic increased profitability and reduced the use of a potent pesticide. The intermingled planting of crop genotypes that have different disease-resistance profiles — called a multiline — can also decrease or even effectively eliminate a pathogen.

Sustainable livestock production

The production of 1 kg of meat can require between 3 and 10 kg of grain. During the past 40 years, global per capita meat production has increased more than 60% (Fig. 3), a trend driven by increasing global per capita incomes, but threatened by stagnant or declining per capita grain production (Fig. 3). Livestock production is becoming an industrial-scale process in which several thousand cattle or pigs, or 100,000 or more chickens, are fed grains and produced in a single facility. In the United States, the average number of animals per livestock operation increased 1.6-fold for cattle, 2.3-fold for pigs, 2.8-fold for egg production and 2.5-fold for broiler chickens over 14 years⁷¹. The average number of pigs per operation increased 2.6-fold from 1990 to 2000 in Canada⁷². Large-scale facilities are economically competitive because of production efficiencies⁷³, but have health and environmental costs that must be better quantified to assess their potential role in sustainable agriculture.

High-density animal production operations can increase livestock disease incidence, the emergence of new, often antibiotic-resistant diseases, and air, groundwater and surface water pollution associated with animal wastes. Current livestock operations are vulnerable to catastrophic loss of animals to disease. For instance, in 1997, an influenza A virus (H5N1) appeared and spread among

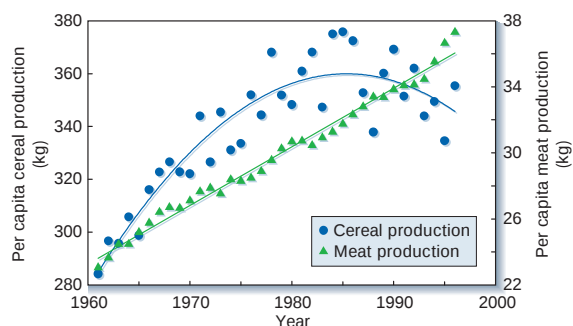


Figure 3 Long-term trends in average per capita food supply. Average annual per capita production of cereal grains and of slaughtered livestock, calculated as total global production for a given year divided by total global population for that year².

Hong Kong chicken-production facilities, killing six humans and leading to the destruction of more than 1.2 million birds. In Britain, foot and mouth outbreaks led to the destruction of 440,000 animals in 1967 and 1.2 million in 2001. Bovine spongiform encephalitis ('mad cow disease') led to the slaughter of 11 million animals in 1996. To help prevent disease associated with high-density facilities, livestock are often fed subtherapeutic doses of the same antibiotics used in human medicines. These prophylactic treatments cause agriculture to use, in total, a larger proportion of global antibiotic production than human medicine¹⁵. Antibiotic-resistant *Salmonella*, *Campylobacter* and *Escherichia coli* strains that are pathogenic to humans are increasingly common in poultry or beef produced in large-scale operations¹⁴.

The handling and disposal of animal wastes are significant problems of high-density animal confinement facilities. Manure lagoons can release high levels of hydrogen sulphide and other toxic gases, volatilize ammonium that greatly increases regional nitrogen deposition, and contaminate surface and ground waters with nutrients, toxins and pathogens. These animal wastes pose health and environmental risks similar to those of human wastes and should be treated accordingly. For example, animal wastes could be treated by composting to create a crop fertilizer that no longer harbours pathogens, and that is applied at appropriate rates and times and with methods that minimize nutrient leaching. This closing of the nutrient cycle decreases dependence on synthetic fertilizer production, and is more efficient when animal and crop production are combined locally.

Pastoral livestock production makes extensive use of ecosystem services and eliminates many of the problems of confinement production. Pastured animals consume plants growing in a field, and plant growth is increased by animal wastes deposited and recycled in the field. Ruminant production on grasslands takes advantage of the high efficiency of ruminant guts to convert low-quality forage into high-protein human foods, including dairy products and beef. When appropriately stocked and managed, grassland-ruminant ecosystems are an efficient, sustainable method of producing high-quality protein with minimal environmental impacts.

Implementing sustainable practices

Farmer incentives are a central issue facing sustainable agriculture. Farmers grow crops or raise livestock to feed their families or to sell and earn a living in a market economy that is becoming increasingly global and competitive. Although some ecosystem services, such as pollination or control of agricultural pests, are of direct benefit to a farmer, other ecosystem services may benefit the public as a whole but be of little or no direct benefit to the farmer. Consider hypoxia in the Gulf of Mexico, which is attributable in large part to nitrogen

runoff from agriculture in the Mississippi River drainage⁴⁵. Reducing nitrogen fertilization on a single farm would (marginally) help reduce hypoxia, but, beyond some point, would also reduce yields and profit. The benefit to the Gulf is of no direct benefit to the farm.

Current incentives favour increased agricultural production at the expense of ecosystem services. Interestingly, many studies indicate that fertilizer-use efficiency could be greatly increased by better matching nutrient inputs to crop demand in time and space^{40,44}, but essential investments in on-farm nutrient-management research and in extension activities that promote such practices have not yet occurred. Similar opportunities for a significant increase in fertilizer efficiency exist for small-scale intensive rice cropping systems in the developing countries of Asia⁷⁴.

How, then, can society accomplish the dual objectives of improving yield levels and food stability and of preserving the quality and quantity of ecosystem services provided by the Earth's land and water resources? Clearly, appropriate incentives are needed. In addition to the practices described in the preceding sections, farmers will need to rely on a rapidly expanding base of biological and agronomic knowledge that is often specific to certain agroecosystems, regions, soil types and slopes. Making the right decisions at the farm level in terms of input-use efficiency, human health and resource protection is becoming an increasingly knowledge-intensive task.

What incentives and policies could lead to the adoption of sustainable farming practices? In 1999, member countries of the Organisation for Economic Co-operation and Development provided US\$283 billion in subsidies to support agricultural production (of which US\$74 billion was for grains)⁷⁵. These funds need to be reoriented to support sustainable practices. Several policy initiatives have tried to level the playing field between agricultural production and production of ecosystem services. A number of countries, including Australia, Canada, European Union (EU) countries, Japan, Norway, Switzerland and the United States, have instituted various forms of 'green payments', that is, payments to farmers who adopt sustainable or environmentally benign farming practices⁷⁶. Norway and Switzerland provide substantial payments for 'landscape maintenance'. The United States' Conservation Reserve Program pays farmers to take land out of production for a specified period, and some countries have also instituted 'environmental cross-compliance' conditions as a prerequisite for farmers to receive agricultural support payments. Other policy options include taxes, removal of subsidies, and implementation of new regulations. A tax on fertilizers or pesticides, or removal of subsidies for these inputs, would discourage excessive use. International policies are needed when actions in one country cause environmental damage in another country, such as for polluted rivers that cross national boundaries, or for emissions of greenhouse gases. But as the negotiations over the Kyoto Protocol on greenhouse gas emissions demonstrate, both the attainment and enforcement of such policies are major challenges.

Consumer incentives are also possible. A broad look at trends in agricultural production shows that many of the elevated environmental impacts projected for the coming 50 years are tied to increased consumption of livestock products and concomitant elevated demand for grains fed to livestock. Pricing and labelling each type of livestock product to reflect the true total costs of its production could provide consumers with important information and with incentives for choosing alternative food products.

Providing the right incentives should help maximize the total return to society of the net benefits of agricultural production. However, many environmental problems and ecosystem services are difficult to monitor and quantify. For nitrogen or pesticide runoff or carbon sequestration, it may be costly to assess environmental performance of individual farms. Rather than basing incentive payments on environmental performance itself, proxies for performance, such as the adoption of certain auditable practices, may be as close as policy can get. The achievement of such objectives will require coordination among federal agencies or ministries for agriculture and for

environment, which often have different objectives. Sustainable agriculture requires addressing the concerns of both groups.

The pursuit of sustainable agriculture will also require substantial increases in knowledge-intensive technologies that enhance scientifically sound decision making at the field level⁷⁷. This can be embedded in physical technology (for example, equipment and crop varieties) or in humans (for example, integrated pest management), but both are essential. However, the challenges of disseminating information on new technologies or on efficient input use and management are enormous, especially in cases where extension programmes are ineffective or completely lacking. The earlier paradigm of science being developed at the international or perhaps national level and then disseminated to farmers should be replaced by an active exchange of information among scientists and farmers. Scientists in developing countries who understand the ecosystems, human culture and demands on local agricultural systems must be actively trained, promoted and brought into the international scientific community.

Substantially greater public and private investments in technology and human resources are needed internationally, especially in low-income nations, to make agricultural systems more sustainable. Global research expenditures are less than 2% of agricultural gross domestic product (GDP) worldwide⁷⁸, being roughly 5.5% of agricultural GDP in developed countries, but less than 1% in developing countries (where most of the increased food demand will occur during the next 50 years). At present, there are few incentives for the private sector to increase investments in lower-income developing countries^{78,79}. Furthermore, unless reward structures also reflect the value of ecosystem services, there will be little incentive for the private sector to invest in sustainable agricultural methods. Without adequate investments, yield gains and environmental protection may be insufficient for a transition to sustainable agriculture.

Implications

The coming 50 years are likely to be the final period of rapidly expanding, global human environmental impacts. Future agricultural practices will shape, perhaps irreversibly, the surface of the Earth, including its species, biogeochemistry and utility to society⁴. Technological advances and current economic forces, including large agricultural subsidies in the United States, EU and Japan, have both increased food availability and decreased the real costs of agricultural commodities during the past 50 years. But the resulting agricultural practices have incurred costs related to environmental degradation, loss of biodiversity, loss of ecosystem services, emergence of pathogens, and the long-term stability of agricultural production.

The goal of sustainable agriculture is to maximize the net benefits that society receives from agricultural production of food and fibre and from ecosystem services. This will require increased crop yields, increased efficiency of nitrogen, phosphorus and water use, ecologically based management practices, judicious use of pesticides and antibiotics, and major changes in some livestock production practices. Advances in the fundamental understanding of agroecology, biogeochemistry and biotechnology that are linked directly to breeding programmes can contribute greatly to sustainability^{6,66}.

Agriculturalists are the *de facto* managers of the most productive lands on Earth. Sustainable agriculture will require that society appropriately rewards ranchers, farmers and other agriculturalists for the production of both food and ecosystem services. One major step would be achieved were agricultural subsidies in the United States, EU and Japan redirected to reward sustainable practices. Ultimately, sustainable agriculture must be a broadly based effort that helps assure equitable, secure, sufficient and stable flows of both food and ecosystem services for the 9,000 million or so people likely to inhabit the Earth. □

doi:10.1038/nature01014

1. Cohen, J. E. *How Many People Can the Earth Support?* (Norton, New York, 1995).
2. Food and Agriculture Organization of the United Nations (FAO). FAO Statistical Databases <<http://apps.fao.org/>> (2001).

3. World Health Organization (WHO). *Public Health Impacts of Pesticides Used in Agriculture* (WHO in collaboration with the United Nations Environment Programme, Geneva, 1990).
4. Tilman, D. *et al.* Forecasting agriculturally driven global environmental change. *Science* **292**, 281–284 (2001).
5. Waggoner, P. E. How much land can ten billion people spare for nature? Does technology make a difference? *Technol. Soc.* **17**, 17–34 (1995).
6. Cassman, K. G. Ecological intensification of cereal production systems: yield potential, soil quality, and precision agriculture. *Proc. Natl Acad. Sci. USA* **96**, 5952–5959 (1999).
7. Cohen, J. E. & Federoff, N. V. *Colloquium on Plants and Population: Is There Time?* (National Academy of Sciences, Washington DC, 1999).
8. Alexandratos, N. World food and agriculture: outlook for the medium and longer term. *Proc. Natl Acad. Sci. USA* **96**, 5908–5914 (1999).
9. Ruttan, V. W. The transition to agricultural sustainability. *Proc. Natl Acad. Sci. USA* **96**, 5960–5967 (1999).
10. Ruttan, V. R. Productivity growth in world agriculture: sources and constraints. *J. Econ. Perspect.* (in the press).
11. Postel, S. *Pillar of Sand: Can the Irrigation Miracle Last?* (Norton, New York, 1999).
12. Vitousek, P. M., Mooney, H. A., Lubchenco, J. & Melillo, J. M. Human domination of earth's ecosystems. *Science* **277**, 494–499 (1997).
13. Carpenter, S. R. *et al.* Nonpoint pollution of surface waters with phosphorus and nitrogen. *Ecol. Applic.* **8**, 559–568 (1998).
14. Smith, K. E. *et al.* Quinolone-resistant *Campylobacter jejuni* infections in Minnesota, 1992–1998. *New Engl. J. Med.* **340**, 1525–1532 (1999).
15. Gorbach, S. L. Antimicrobial use in animal feed—time to stop. *New Engl. J. Med.* **345**, 1202–1203 (2001).
16. Cassman, K. G. & Pingali, P. L. Intensification of irrigated rice systems: learning from the past to meet future challenges. *Geoljournal* **35**, 299–305 (1995).
17. Daily, G. C. *Nature's Services: Societal Dependence on Natural Ecosystems* (Island, Washington DC, 1997).
18. National Research Council. *Nature's Numbers: Expanding the National Economic Accounts to Include the Environment* (National Academy Press, Washington DC, 1999).
19. Daily, G. C. *et al.* The value of nature and the nature of value. *Science* **289**, 395–396 (2000).
20. Nye, P. H. & Greenland, D. J. *The Soil Under Shifting Cultivation* Tech. Commun. No. 51 (Commonwealth Agricultural Bureau of Soils, Harpenden, 1960).
21. Hector, A. B. *et al.* Plant diversity and productivity experiments in European grasslands. *Science* **286**, 1123–1127 (1999).
22. Tilman, D. *et al.* Diversity and productivity in a long-term grassland experiment. *Science* **294**, 843–845 (2001).
23. Loreau, M. S. *et al.* Biodiversity and ecosystem functioning: current knowledge and future challenges. *Science* **294**, 804–808 (2001).
24. Power, A. G. Linking ecological sustainability and world food needs. *Dev. Sustainability* **1**, 185–196 (1999).
25. Manning, R. *Food's Frontier: The Next Green Revolution* (North Point, New York, 2000).
26. Kates, R. W. Ending hunger: current status and future prospects. *Consequences* **2**, 3–12 (1996).
27. Sen, A. *Poverty and Famines: An Essay on Entitlement and Deprivation* (Clarendon, Oxford, 1981).
28. Plucknett, D. L. International agricultural-research for the next century. *Bioscience* **43**, 432–440 (1993).
29. Conway, G. *The Doubly Green Revolution: Food for All in the Twenty-First Century* (Penguin, London, 1997).
30. Barnett, V., Payne, R. & Steiner, R. *Agricultural Sustainability: Economic, Environmental and Statistical Considerations* (Wiley, Chichester, 1995).
31. Young, A. Is there really spare land? A critique of estimates of available cultivable land in developing countries. *Environ. Dev. Sustainability* **1**, 3–18 (1999).
32. Cassman, K. G. in *Crop Science—Prospects and Progress* (eds Nosberger, J., Geiger, H. H. & Struik, P. C.) 33–51 (CAB International, Wallingford, 2001).
33. Cassman, K. G. & Dobermann, A. in *Rice Research and Production in the 21st Century: Symposium Honoring Robert F. Chandler, Jr.* (ed. Rockwood, W. G.) 79–100 (International Rice Research Institute, Los Banos, Philippines, 2001).
34. Reynolds, M. P., Rajaram, S. & Sayre, K. D. Physiological and genetic changes of irrigated wheat in the post-green revolution period and approaches for meeting projected global demand. *Crop Sci.* **39**, 1611–1621 (1999).
35. Peng, S., Cassman, K. G., Virmani, S. S., Sheehy, J. & Khush, G. S. Yield potential trends of tropical rice since the release of IR8 and the challenge of increasing rice yield potential. *Crop Sci.* **39**, 1552–1559 (1999).
36. Duvick, D. N. & Cassman, K. G. Post-green-revolution trends in yield potential of temperate maize in the north-central United States. *Crop Sci.* **39**, 1622–1630 (1999).
37. Pinstrup-Andersen, P. & Pandya-Lorch, R. Food for all in 2020—can the world be fed without damaging the environment. *Environ. Conserv.* **23**, 226–234 (1996).
38. Vitousek, P. M. & Matson, P. A. in *The Biogeochemistry of Global Change: Radiative Trace Gases* (ed. Oremland, R. S.) 193–208 (Chapman and Hall, New York, 1993).
39. Galloway, J. N., Levy, H. II & Kashibhatla, P. S. Year 2020: consequences of population growth and development on deposition of oxidized nitrogen. *AMBIO* **23**, 120–123 (1994).
40. Cassman, K. G., Dobermann, A. & Walters, D. Agroecosystems, nitrogen-use efficiency, and nitrogen management. *AMBIO* (in the press).
41. Smil, V. Nitrogen in crop production: an account of global flows. *Global Biogeochem. Cycl.* **13**, 647–662 (1999).
42. Smil, V. Phosphorus in the environment: natural flows and human interferences. *Annu. Rev. Energy Environ.* **25**, 53–88 (2000).
43. Howarth, R. W. *et al.* Regional nitrogen budgets and riverine N & P fluxes for the drainages to the North Atlantic Ocean: natural and human influences. *Biogeochemistry* **35**, 75–139 (1996).
44. Matson, P. A., Naylor, R. & Ortiz-Monasterio, I. Integration of environmental, agronomic, and economic aspects of fertilizer management. *Science* **280**, 112–115 (1998).
45. Downing, J. A. *et al.* *Gulf of Mexico Hypoxia: Land and Sea Interactions* Task Force Report No. 134 (Council for Agricultural Science and Technology, Ames, IA, 1999).
46. Cicerone, R. J. & Oremland, R. S. Biogeochemical aspects of atmospheric methane. *Global Biogeochem. Cycl.* **2**, 299–327 (1988).
47. Hall, S. J., Matson, P. A. & Roth, P. NO_x emission from soil: implications for air quality modeling in agricultural regions. *Annu. Rev. Energy Environ.* **21**, 311–346 (1996).

48. Delmas, R., Serca, D. & Jambert, C. Global inventory of NO_x sources. *Nutr. Cycl. Agroecosyst.* **48**, 51–60 (1997).
49. Chameides, W. L., Kasibhatla, P. S., Yienger, J. & Levy, H. Growth of continental-scale metro-agroplexes, regional ozone pollution, and world food production. *Science* **264**, 74–77 (1994).
50. Vitousek, P. M. *et al.* Human alteration of the global nitrogen cycle: sources and consequences. *Ecol. Applic.* **7**, 737–750 (1997).
51. Prather, M. D. *et al.* in *Climate Change 2001: The Scientific Basis* (Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change) (eds Houghton, J. T. *et al.*) 239–287 (Cambridge University Press, Cambridge, 2001).
52. Frink, C. R., Waggoner, P. E. & Ausubel, J. H. Nitrogen fertilizer: retrospect and prospect. *Proc. Natl Acad. Sci. USA* **96**, 1175–1180 (1999).
53. Drinkwater, L. E., Wagoner, P. & Sarrañtonio, M. Legume-based cropping systems have reduced carbon and nitrogen losses. *Nature* **396**, 262–265 (1998).
54. Matson, P. A., Billow, C. & Hall, S. Fertilization practices and soil variations control nitrogen oxide emissions from tropical sugar cane. *J. Geophys. Res.* **101**, 18533–18545 (1996).
55. Cassman, K. G., Kropff, M. J., Gaunt, J. & Peng, S. Nitrogen use efficiency of irrigated rice: What are the key constraints? *Plant Soil* **155/156**, 359–362 (1993).
56. Peng, S. *et al.* Increased N-use efficiency using a chlorophyll meter on high-yielding irrigated rice. *Field Crops Res.* **47**, 243–252 (1996).
57. Woerner, P. L. & Swift, M. J. *The Biological Management of Tropical Soil Fertility* (Wiley, Chichester, 1994).
58. Roberston, G. P. in *Ecology in Agriculture* (ed. Jackson, L. E.) 347–365 (Academic, San Diego, 1997).
59. Gleick, P. Water and conflict: fresh water resources and international security. *Int. Security* **18**, 79–112 (1993).
60. Postel, S. L., Daily, G. C. & Ehrlich, P. R. Human appropriation of renewable fresh water. *Science* **271**, 785–788 (1996).
61. Dynesius, M. & Nilsson, C. Fragmentation and flow regulation of river systems in the northern third of the world. *Science* **266**, 753–762 (1994).
62. Seckler, D., Barker, R. & Amarasinghe, U. Water scarcity in the twenty-first century. *Int. J. Water Resources Dev.* **15**, 29–42 (1999).
63. Postel, S. L. *Last Oasis: Facing Water Scarcity* (Norton, New York, 1992).
64. Naylor, R. Energy and resource constraints on intensive agricultural production. *Annu. Rev. Energy Environ.* **21**, 99–123 (1996).
65. Charles, D. Seeds of discontent. *Science* **294**, 772–775 (2001).
66. DeVries, J. & Toennissen, G. *Securing the Harvest: Biotechnology, Breeding, and Seed Systems for African Crops* (CAB International, Wallingford, 2001).
67. Oldeman, L. R. in *Soil Resilience and Sustainable Land Use* (eds Greenland, D. J. & Szabolcs, J.) 99–118 (CAB International, Wallingford, 1994).
68. Ortiz, R. Critical role of plant biotechnology for the genetic improvement of food crops: perspectives for the next millennium. *J. Biotechnol.* **1**, 1–8 (1998).
69. Palumbi, S. R. Humans as the world's greatest evolutionary force. *Science* **293**, 1786–1790 (2001).
70. Zhu, Y. *et al.* Genetic diversity and disease control in rice. *Nature* **406**, 718–722 (2000).
71. United States General Accounting Office (GAO). *Animal Agriculture: Information on Waste Management and Water Quality Issues* GAO/RCED-95-200BR (GAO, Washington DC, 1995).
72. Statistics Canada. Number of farms reporting pigs and average number of pigs per farm. *Livestock Statistics Cat. No. 23-603-UPE* <<http://www.statcan.ca>> (2002).
73. Martin, L. Costs of production of market hogs. *Western Hog J.* (Banff Pork Semin. 2000 Spec. Edn) 24 (2000).
74. Dobermann, A. *et al.* Site-specific nutrient management for intensive rice cropping systems in Asia. *Field Crops Res.* **74**, 37–66 (2002).
75. Organisation for Economic Co-operation and Development (OECD). *Agricultural Policies in OECD Countries: Monitoring and Evaluation 2000* (OECD, Paris, 2000).
76. Organisation for Economic Co-operation and Development (OECD). *Agricultural Policies in OECD Countries: Monitoring and Evaluation* (OECD, Paris, 2001).
77. Byerlee, D. Modern varieties, productivity, and sustainability—recent experience and emerging challenges. *World Dev.* **24**, 697–718 (1996).
78. Pardey, P. G. & Beintema, N. M. *Slow Magic: Agricultural R&D a Century after Mendel* (International Food Policy Research Institute, Washington DC, 2001).
79. Falcon, W. & Fowler, C. Carving up the commons. *Food Policy* (in the press).

Acknowledgements

We thank the National Science Foundation for support and N. Larson for assistance.

Enhancing the crops to feed the poor

Jikun Huang*, Carl Pray† & Scott Rozelle‡

*Center for Chinese Agricultural Policy, Institute of Geographic Sciences and Natural Resource Research, Chinese Academy of Sciences, 917 Building, Anwai Datun Road, Beijing 100101, China

†Department of Agricultural, Food and Resource Economics, Rutgers University, 55 Dudley Road, New Brunswick, New Jersey 08901-8520, USA

‡Department of Agricultural and Resource Economics, University of California, 1 Shields Avenue, Davis, California 95616, USA (e-mail: rozelle@primal.ucdavis.edu)

Solutions to the problem of how the developing world will meet its future food needs are broader than producing more food, although the successes of the 'Green Revolution' demonstrate the importance of technology in generating the growth in food output in the past. Despite these successes, the world still faces continuing vulnerability to food shortages. Given the necessary funding, it seems likely that conventional crop breeding, as well as emerging technologies based on molecular biology, genetic engineering and natural resource management, will continue to improve productivity in the coming decades.

Billions of people struggle for a better life in the developing world, but they are able to improve their prospects of achieving this only when there is abundant and affordable food available¹. Food security for the poor is dependent on issues such as access to the resources needed to buy or produce their own food²; nevertheless, welfare increased dramatically for many after the Second World War, in part because of the huge increase in agriculture's ability to produce food. Improving quality of life in the twenty-first century will likewise require as much, if not more, effort in increasing global food production. One of the great challenges of the coming decades will be to produce the food and fibre that is needed to feed and clothe those in the poorer parts of the world. And although from some perspectives this seems like an impossible task — in the same way that it must have to the doomsday forecasters since the days of Malthus — there are many reasons to believe it can be achieved.

In this review, we explore how technology can help the developing world meet its food needs in the twenty-first century. We begin by discussing the role of technology in generating past growth in productivity and output by analysing the successes and failures of the Green Revolution. Despite the past successes, the world's continuing vulnerability to food shortages is illustrated. The constraints that are holding back food production are examined, and we divide these into those that can be addressed by traditional crop breeding and agronomic techniques, and those that can be best solved by biotechnology and other high-technology approaches. We then shift our focus to the future. Drawing on a survey of prominent scientists and research administrators in China and interviews with scientists elsewhere in the world, we assess the technologies that are currently available and those that hold promise in the future. Finally, we turn our attention to who will create the new technologies and where the resources to create them will come from.

Past achievements and persistent vulnerability

With the exception of several short periods (for example, the mid-1970s), world food production has expanded continuously since the 1960s. And, while the industrialized countries have contributed significantly to the world's food supply, the developing world also has played an important role and has been a major beneficiary. From a world facing the prospects of severe global famine, the input-responsive

plant varieties of the Green Revolution, together with subsequent investments in water control, intensification of chemical input use and further genetic discoveries, raised food production to levels that no one would have dared predict³. Global yields rose by 2.42% annually between 1961 and 1966 (Fig. 1). With the exception of Africa, farmers in developing and developed countries nearly doubled their per-hectare output of cereal production, increasing yields during this time by 3.16% annually. Sown area, although actually trending down slightly worldwide (−0.04% per year), still expanded annually by about 0.25% in developing countries. With more food available, falling food prices, increased food trade and rising consumption led the way to lower malnutrition and falling poverty rates in many parts of the world⁴.

Past success, however, does not guarantee a food-abundant world in the coming decades. Growth rates of yields have slowed during the period between 1987 and 2001 (Fig. 1). Moreover, the demographic pressures in the twenty-first century will be unprecedented. The world's population will reach 8 billion by 2025. By 2020, increasingly wealthy and urbanized consumers and the 2 billion new mouths will demand 40% more food⁵. Rosegrant *et al.* estimate that food and feed production must continue to rise annually by 1.2% to satisfy the demand of the world's population by 2020 (ref. 6).

Increasing yields

Although it may seem easy to reach growth rates of food production that are approximately half the average of the past 40 years, the exhaustion of some past sources of growth makes future yield expansion as great a challenge as in the past⁷. Populations have encroached on almost all of the world's frontiers, leaving little new land that is cultivatable with current technologies. Widespread adoption of modern varieties and intensive use of inorganic fertilizers and chemical pesticides have pushed yields in many areas of developing countries to levels that rival those reached on farms in developed countries and on the experimental fields of agronomists. For example, nearly 100% of farmers in China use improved varieties of rice, wheat and maize⁸, and producers in south Asia and Latin America use modern cereal varieties on more than 80% of their sown area⁹. The gap between the actual yields of farmers and those that are attainable on experimental plots, given the current resource base and economic environment, has narrowed in a

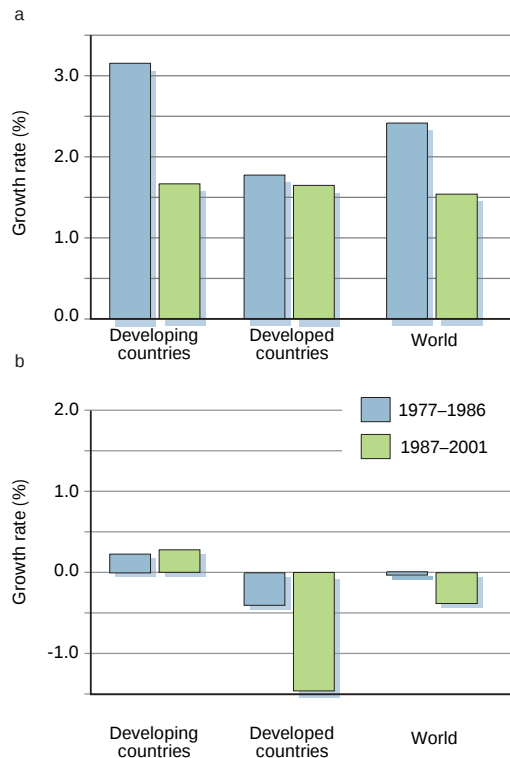


Figure 1 Annual growth rate of cereal yields and sown area in developing and developed countries, 1977–2001. Data from Food and Agriculture Organization of the United Nations.

number of the world's developing nations⁷. Whereas in the past, large increases in output per hectare could come from transferring existing technology from other parts of the world or from increasing the use of fertilizer, water and other inputs, in many parts of the world these gains have been largely exhausted.

Pressure on the environment and resource base

Environmental pressures also may threaten the sustainability of the world's food productivity gains. Some scientists have voiced concerns that the intensification of farming systems may not be sustainable because of systematic degradation of the resource base and environment^{10,11}. The degradation, which is reflected in declining yields in long-term agronomic experiments and decelerating aggregate yields at the national level, appears in many intensive-cropping systems in different areas of the world and is attributable to a number of factors^{12,13}. Scientists at the International Rice Research Institute (IRRI) hypothesize that rice yields fall over time owing to a decline in soil nitrogen in water-saturated soils, increased incidence of disease with high nitrogen use, and a build up of soil pests because of continuous monocropping¹⁴. In the highly productive Punjab region of India and Pakistan, the productivity of wheat–rice systems is declining because of falling concentrations of organic matter and phosphorus in the soil and increasing salinity in the groundwater^{15,16}. At the other extreme, African farmers, who use almost no modern inputs, mine the soil's natural nutrient base by adoption of continuous cropping without the use of modern inputs and elimination of traditional fallowing practices¹⁷.

An arms race between pests and technology is being played out with rising disease and insect pressures threatening to reduce the gains provided by modern varieties. Because the use of modern varieties allowed

intensified cultivation within a season and over the year, a rise in the amount of plant matter per hectare and the shortening of the length of time between plantings, especially in the case of rice, increased the incidence of diseases and the size of insect populations¹⁸. Farmers fought back with chemical pesticides. Breeders reinforced the pest-control efforts of farmers by producing a second generation of pest-resistant modern varieties¹⁹. But although the chemical attack on pests and intense use of resistant varieties was effective in reducing pest-related damage, it triggered a number of unexpected negative effects²⁰. First, in the process of killing destructive pests, pesticides also invariably destroy the predators of these pests. Second, natural selection is creating a class of pests that is becoming resistant to many of the most popular and relatively safe pesticides. It has been shown that in some developing countries, the cost of the adverse health consequences for the farmer applying the pesticide more than offsets the savings that the farmer earns by reducing the loss of pest-inflicted damage to the crop²¹. These problems, responses and subsequent reactions exemplify the delicate nature of science's long-term fight to raise enough food for a hungry and burgeoning population.

Finally, increasingly severe water shortages may constrain future yield rises. In the past, irrigation has been crucial in increasing crop yields and facilitating the shift to new technologies and more intense cultivation. India, for example, has invested more in water control than any other activity²². China invests more than ten times more into water control than into agricultural research²³. Huang *et al.* illustrate that yields and income, including those of the poor, are raised by between 30 and 100% when cultivated area is irrigated²⁴. However, by some accounts, water is becoming the most constraining input to agriculture in many nations²⁵. Up to one-half of the world's population lives in a water-scarce environment. In other areas, lack of infrastructure allows much of the water that comes from natural sources, such as rainfall, to run off without being able to be used. Additionally, city users and industries will undoubtedly out-compete agriculture for fixed amounts of water resources in many regions.

Breaking productivity constraints

The twenty-first century needs another Green Revolution to elevate global food production. Because of the limited amount of land and water in many parts of the world, the only way to expand production is by developing a technology that increases output per unit of input. In the same way that fertilizer-responsive dwarf varieties and hybrid cultivars increased the productivity of land and other inputs after 1950, new technologies are needed to create additional productivity gains in the coming years. Some of these gains will be similar to those of the past. However, unlike the first two stages of the Green Revolution, which centred initially on favourable, irrigated areas of the world in the 1960s and 1970s, and then on more favourable (in terms of soil and other growing conditions) rain-fed areas in the 1980s and 1990s, the next generation of even more powerful technologies may be able to address a wider array of constraint sets that can allow production to rise. The job of scientists, however, will be made more complicated, as different technologies will be needed to address different sets of questions in different areas of the world.

The need to raise yield plateau

Because much of the exploitable yield gap between the farm and experimental station has been eliminated in many of the world's favourable areas, new technologies for these regions almost certainly will depend on increased investment in research on advanced breeding techniques, crop physiology and molecular biology⁷. Virmani *et al.* believe extension of hybrid-variety development to rice, wheat and other crops can increase yields by 20% (ref. 26). The spread of hybrid rice in China and the associated rise in yields demonstrates the validity of this approach²⁷. China's plant breeders have been working on a new approach to producing hybrid rice, which will allow easier and cheaper seed production²⁸. Plant physiologists have also created a model for a new plant type that could potentially increase rice yields



Figure 2 Comparison of traditional, improved and ideal rice plant types. Reproduced from ref. 29, with permission.

by 30% (to nearly 15 tonnes per hectare, if it is combined with hybrid-variety development) by increasing the plant's inherent production efficiency²⁹. IRRI scientists report that farmers in China's subtropical provinces are already using the first generation of this new technology — after several years of testing on experimental stations in the Philippines and Indonesia, new plant-type lines that yield 20% more than traditional varieties in tropical areas have moved into seed board trials (Fig. 2).

Narrowing the yield gap

Byerlee *et al.* identify a distinct set of constraints that are holding back productivity gains in favoured areas with significant exploitable potential (referred to hereafter as semi-favourable, rain-fed areas)⁷. Whereas the yield gaps in the most favourable areas are narrow, there are still large areas of the world in which output is only one-third or less of those in local experiment stations. Farmers in these areas, which are largely in the rain-fed belts of Latin America, south Asia and Africa, currently use low levels of inputs and are poorly trained in modern farming techniques. Moisture in these areas, although variable, is sufficient for producing high yields in most years. But the crops of farmers in these areas are frequently nutrient-starved and must be planted in soils that are relatively fragile. It is not that the nutrients are absent; more often it is the case that they are unevenly spread across the landscape and frequently out of balance in terms of their nitrogen–phosphate–potassium mix and/or lack micro-nutrients.

Unlike the most favourable areas, semi-favourable, rain-fed areas need adaptive plant breeding and extension of agronomic practices and farm management techniques to overcome some of their largest constraints. Scientists at the West Africa Rice Development Association crossed Asian rice varieties with African varieties to develop new varieties called NERICA (for New Rice for Africa). These varieties combine the weed-control and drought-resistant characteristics of their African parent with the high-yielding characteristics of their Asian parent, and are now spreading rapidly in West Africa. In some case, new technologies that are simple and could simultaneously relax several constraints are possible. For example, the new rice varieties being bred with higher levels of iron and zinc to alleviate malnutrition may also increase yields in iron- and zinc-deficient soils. However, education levels, literacy and cash availability are so low and access to communications and transportation are so poor that infrastructure investments are often preconditions for overcoming the agricultural constraints in these areas.

Marginal areas, as their name implies, include those regions where yields are severely constrained by climatic stress (for example, drought), fragile soils and non-existent or non-functional infrastructure. During the last half of the twentieth century, scientists

directed less of their work towards these areas³⁰. Agricultural technology had less to offer, as modern varieties during this period responded to water and fertilizer; instead, some of the most productive work was done in the area of natural resource management. People in these regions did benefit from the Green Revolution, albeit indirectly, by gaining access to lower-cost foods in markets and migrating to areas where they could hire themselves out as wage earners on the farms that had adopted the labour-using new technologies.

Many of the world's poorest people live in these marginal areas, and they lack the purchasing power or labour skills to reap the indirect benefits of agricultural technologies introduced in favourable or semi-favourable areas. Because there was less work done on the problems of poor areas during the last Green Revolution, there is almost certainly much more scope for improving the livelihood of those who live there now. To ensure that people in these areas benefit from the development of the next generation of Green Revolution technologies, scientists and policy-makers need to make sure the investments in the human and physical capital are made so that farmers are able to utilize them when they become available.

Genetic engineering and other new technological approaches

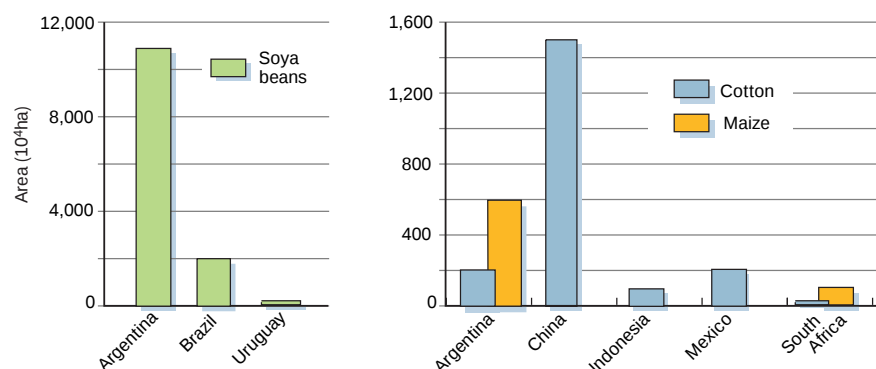
In addition to the research from traditional agricultural science, plant biotechnology research — which includes genetic engineering and the transfer of genes from unrelated plants and microorganisms — started to show important benefits for farmers in developed countries in the mid-1990s and in favourable regions of developing countries in 1997 (ref. 31). Farmers in both the developed and developing world most frequently adopt Roundup Ready soya beans, *Bt* cotton (containing a gene for an insecticide derived from the bacterium *Bacillus thuringiensis*), and *Bt* yellow maize or corn. In the developing world, genetically modified (GM) crops have had the largest impact in Argentina, Brazil (where they are still not officially approved for use, but have been purchased from Argentina), China and South Africa (Fig. 3). James estimates that approximately 5.5 million farmers in developing countries currently receive direct benefits of GM technologies³². However, not all farmers in developing countries have access to the new technologies. Almost all the GM crops currently being grown in developing countries are temperate-region cash crops, and few are being used for food (including yellow maize, a crop used primarily as animal feed). Almost all the new genes that are in commercially grown GM crops, except for those in China's *Bt* cotton crop (which were produced by its own domestic scientists), were produced by Monsanto, a large multinational life science company, based in the United States.

Since the development of GM crops in the 1980s and their release in the mid-1990s, investments have focused on a narrow range of applications and adoption has been mainly by producers in industrialized countries. For example, half of all experimental trials in developed countries have been on varieties that have been made to be insect resistant or herbicide tolerant, or both³¹. More than two-thirds of trials in developing countries work on these same traits (Fig. 4). Because of the nature of the new technologies (which allow the private-company developers to capture part of the rent of the new technology through the sale of hybrid seeds or through sale of the chemical herbicide that must be used with the herbicide-tolerant varieties), the most enthusiastic adopters have been farmers in developed countries. Adoption in developed countries (40 million hectares), where most farmers use relatively less labour and capital is less of a constraint, exceeds that in developing countries (13 million hectares) by more than 300%. Moreover, except for the case of small cotton farmers in China, Mexico and South Africa (who adopted GM technologies on a measurable scale only in the late 1990s), adoption in developed countries and in Argentina (an exception in the developing world) is almost exclusively by commercial producers.

Benefits and limitations of genetic engineering

GM technologies have benefited the farmers who have adopted them, mainly through time-saving gains, increased yields and reduced

Figure 3 Area sown to genetically modified crops in developing countries in 2001. Data from ref. 31; except Brazil which is from authors' interviews.



chemical pesticide inputs. Herbicide-resistant soya beans in Argentina have reduced costs of production per hectare through a reduction in herbicide applications³³. The average *Bt* cotton farmer in China has reduced pesticide sprayings for the Asian boll worm from 20 to 6 times per year and produces a kilogram of cotton for 28% less cost than the farmer using non-*Bt* varieties³⁴. Mexican and South African *Bt* cotton farmers increased the yields at the same time that they reduced their costs^{35,36}. The reduction in pesticide use not only saves farmers the financial outlay for insecticides, but also reduced the incidence of insecticide poisonings³⁷.

Although the potential exists in the future for increasing food production and alleviating constraints on cereal production in semi-favourable and marginal areas of developing countries, progress so far is limited. No GM varieties of a major food grain are currently being grown in developing countries, and there is very little work being done on crops grown in many marginal areas, such as millet, cassava or beans. But field trials for bio-safety clearance of GM varieties show that some major GM food crops are in the pipeline, and a few countries are actually releasing or close to releasing GM food crops. China's scientists, for example, are working on GM rice, potato and peanuts, crops that have been largely ignored in the developed world. Researchers in other developing countries are working on sugarcane, papaya and a number of other tropical crops. South Africa is leading the way in growing GM subsistence crops with the production of GM white maize, the first harvest of which will take place this year³⁸. Other major food crops that are in the final stages of testing before commercial release are *Bt* rice, disease-resistant rice and *Bt* maize in China, and virus-resistant sweet potato and *Bt* maize in Kenya^{32,34}. Much more than in developed countries, biosafety is emerging as a principal constraint on release of GM organisms in developing countries.

Like developed countries, the characteristics of GM crops that are in the pipeline in developing countries are overwhelmingly focused on herbicide tolerance and insect resistance. Except for China, 80% of the field trials are on varieties that contain these characteristics individually or 'stacked' together³⁹. Field trials of crops that were being promoted primarily for higher yield were less than 1% of the field trials in developing countries. In China, however, scientists are experimenting with nutrition-enhanced varieties of rice, shelf-life-enhanced varieties of tomatoes, and other characteristics.

A role for other technologies

In addition to high-technology solutions, there is almost certainly a role for further research in other areas such as natural resource management and post-harvest handling and processing. Scientists working in natural resource management have developed technologies such as zero tillage and integrated pest management that have led to significant gains in some countries. The adoption of zero tillage in South America on 26 million hectares can be considered a Green

Revolution in itself⁴⁰, and adoption of such systems is now occurring rapidly in south Asian rice–wheat systems. Eliminating tilling can cut production costs by 50%, save labour, and have a positive impact on the environment by reducing erosion, the volume and toxicity of agrochemicals, and tractor fuel consumption. Crop rotations that can support high yields without creating adverse environmental effects also show promise in new research¹⁶. Additionally, there may also be room for economic improvements in the way that food moves from the farm to the consumer, eliminating some of the waste that currently occurs.

Future food technology

Although in a fertile imagination there is no limit to the increases that science might deliver over the coming decades, it is more prudent to rely on leading scientists to assess what is possible in the foreseeable future. There is considerable debate about whether conventional plant breeding can continue to generate yield increases and provide farmers with ways of reducing input constraints. Although recent attention has focused on the products produced by plant biotechnology, conventional plant breeding has contributed much more to yield increases than has biotechnology. Despite bold promises, the application of molecular biology and knowledge-intensive technologies has been limited to a small number of traits in a limited number of crops. Only a few multinational, private life-science companies have delivered their new genetic technologies to the market place. Likewise, recent developments in computer, satellite and mechanical

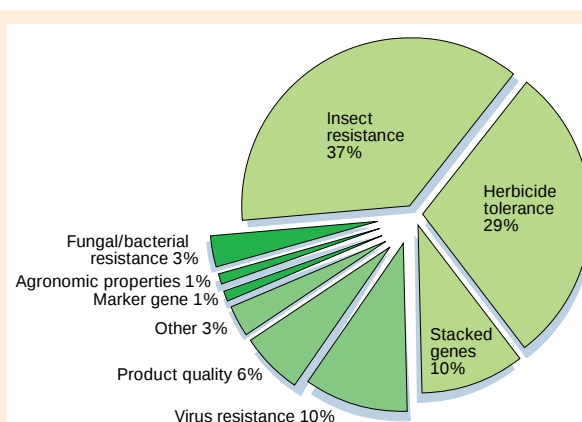


Figure 4 Genetically modified crop traits tested in developed countries, 1987–2000. Data from ref. 31.

Table 1 Crop technology breakthrough predictions in 2020 and effects on crop production in China

Technology	Expected effects on crop production (per cent of those expecting a breakthrough in the technology)				
	Likelihood of significant technological progress	Increases in yields and efficiency	Allows increase in sown area	Allows increase in cropping intensity	Other effects (e.g. on nutrition or on pollution)
Conventional plant breeding	97	49	17	10	24
Genetic engineering	97	43	23	6	28
Precision technology	71	56	11	15	18
Information technology	76	41	8	13	38
For use in water-short areas	97	49	31	10	10
Labour-saving technologies	76	53	16	14	18
For increasing nutrition	82	30	2	2	66
For sustaining productivity	85	54	14	16	16

Source: Author's survey.

equipment interfaces have met with only limited success in developed countries and are almost non-existent in poorer nations⁴¹.

China's version of the future of technology

But many scientists, from both developing and developed countries, believe that recent achievements in plant biotechnology, precision agriculture and other new research areas are only the tip of the iceberg. A survey and interviews about the future potential of technologies show that agricultural scientists predict robust technological growth in the next 20 years. Most systematically, from a survey mailed to leading scientists in China's agricultural research system, 34 respondents gave their prediction about the types of future technologies and the nature of the impacts on producers. Although scientists in China may not be representative of the rest of the developing world, given the large size and successful past of China's research system in producing new technologies from conventional and non-conventional technologies, examining the opinions of many of the nation's leading researchers may be instructive in what future path the developing world will take.

Based on the responses of scientists from 25 research institutes and university departments in China, all but one scientist (97%) predicted important discoveries in the areas of both conventional crop breeding and genetic engineering (Table 1, rows 1 and 2). China's scientists, who were the first to create and commercially extend hybrid rice, believe they can improve the quality of hybrid rice varieties to enable them to tap the hybrid vigour that can provide up to 15–20% higher yields. Breeders are also developing hybrids for wheat, soya beans, rapeseed and other crops. Plant biotechnologists, who already have created an impressive array of GM technologies, believe that they have the potential for using transformation, marker-assisted selection and tissue culture to produce a new generation of technology (Box 1).

Although China's GM crops include those that are insect resistant, scientists work on a much broader range of products, in part because they are nearly all employed in the public sector. As such, they can undertake research on crop technologies that may be difficult to protect from the perspective of intellectual property rights. Interestingly, whereas scientists predict that the main effect of future technologies from traditional crop breeding and genetic engineering will be increasing yields, the second largest effect will be to increase the nutritional value of cereals and reduce the use of inputs that are causing environmental and health concerns. The complementarity between biotechnology and traditional crop breeding, at least in a developing country such as China, also means that the nation needs to invest in both types of programme.

Nearly all of the scientists surveyed (97%) also believed that water-saving technologies will affect agricultural production in the coming 20 years (Table 1, row 5). Breeders and molecular biologists forecast that drought-tolerant varieties would become increasingly

available. Identification of genes that can combat abiotic stresses, such as short-term drought, should enable scientists to create varieties that can greatly improve yields in the favourable rain-fed areas. Agronomists foresee new types of plastics that can be used in a profitable and sustainable way for capturing soil moisture and irrigating crops more efficiently. Genetic and agronomic approaches to saving water may also be able to slow or halt the long-term trend of China's falling sown area. Because many of the new marginal areas are in relatively poor regions of the country, the new technologies will have positive, indirect effects on efforts to alleviate poverty.

A lower percentage of China's scientists believe that precision agriculture (71%) and information technology (76%) can increase the performance of agriculture during the next 20 years (Table 1, rows 3 and 4). The most frequently cited technology is the use of satellite communications to provide extension services to farmers, even in remote villages. Several scientists also believe strongly that easy-to-use electronic tools, such as soil nutrient detectors, could increase efficiency and expand yields without the high human capital requirements that are currently demanded by most computer-based precision agriculture technologies. A large fraction of the surveyed scientists also thought that labour-saving, nutrition-enhancing and environmentally friendly technologies were promising in terms of their ability to raise yields and help the poor (Table 1, rows 6 to 8).

Debate on the future of crop breeding and other programmes

Interviews with scientists from the United States, India and the international agricultural research community mostly support the findings of the China survey. Interviews with 22 leading scientists and observers of international research indicate there is still the potential to increase productivity by conventional breeding, in the form of both output per unit of land and saving on conventional inputs. In addition, these scientists believe that biotechnology will have a major impact on yield per hectare and cost of production by 2010 and will continue thereafter. They also predicted that there is scope for important impacts on water saving and labour saving technology.

Some comments, however, demonstrate differences in perceptions between scientists in developing and developed countries. One director of a large plant-breeding programme in a major international centre (that had both plant breeders and biotechnologists) believed that the prospects for further technological discoveries by conventional crop breeders would last only 10 more years. After that, most additional progress will come from biotechnology. The difference with the perception of China's scientists, who saw a long-term role for plant breeding, may be one of individual opinions. Alternatively, it may reflect the fact that in developing countries the labour costs for hiring a plant breeder are much lower and affordable than many of the capital investments that are needed to set up and maintain modern biotechnology programmes. Others have

found that although the economic returns to crop breeding are linear, costs are exponential⁷.

Making the investments to create the technologies

The need for new technology and the promise that it holds challenges nations in developed and developing countries to create a set of institutions that can deliver the technologies that scientists foresee as possible. After growing rapidly during the post-Second World War period, investments in agricultural research have on average decelerated, although there are differences among the regions of the world⁴². In the period 1991–1996, public agricultural research expenditures climbed at 3.6% annually in developing countries⁴³, compared with an annual rate increase of 0.2% in developed nations. Unfortunately, sub-Saharan Africa, the region that may need the research most, suffered a decline in agricultural research during the 1990s. By the mid-1990s, developed and developing countries spent US\$33 billion per year on agricultural research, about 1.04% of the value of agricultural output. The economic rate of return to investments in agricultural research are regularly calculated to be very high, even though there are almost always lags between the time the investment is made and the time that positive returns are earned⁴⁴.

Public investments

Because many agricultural research investments are public goods, governments are responsible for making them. Public investment will almost certainly continue to be one of the main sources of funding in the coming years. Governments accounted for about two-thirds of agricultural research spending in the mid-1990s (about US\$22 billion)⁴³. But despite this large total, and the high returns to their investments, agricultural research in many countries is in crisis. Funding levels are insufficient to generate a steady flow of technology. Poor incentives in large public bureaucracies often dampen the effectiveness of public research organizations⁴⁵.

In developing countries, the public sector dominates research expenditures in traditional agricultural research fields and in biotechnology. Although funding of traditional fields has waned, the promises of basic biology and biotechnology have induced leaders to increase public spending on agricultural research. For example, the governments of China³⁴, Brazil, India⁴⁶ and South Africa³⁸ have each made major investments in agricultural biotechnology research in the past decade. Because of these investments, many basic scientists from research institutes and universities outside the traditional agricultural research system have begun working on agriculture-related topics.

Role of the private sector

The private sector also is assuming a larger role, especially in biotechnology research and development in developing countries. In the industrialized nations, private firms contribute more than half of all the agricultural research and fund most of the biotech research and technology development. Private firms are growing rapidly in developing countries, although at present they contribute about only about 5% (ref. 43). Despite the relatively small investments in agricultural research, private firms are the main source of plant biotechnology contributions in developing countries, accounting for at least 70% of the field trials of GM plant varieties³⁹.

In developing countries, the private sector will continue to have a limited role in research and technology transfer and the public sector faces a period of scarce funding⁴⁷. As a result, the public sector agricultural research systems need to have a carefully defined agenda in order to reduce hunger. Because of the importance of spillovers in agricultural research (that is, because the research output from one region will often be able to be used by those in other regions), and because of the strong economies of scale that are present in plant breeding⁴⁸, a supra-national organization is needed to increase investment internationally. Traditionally, the Consultative Group on International Agricultural Research (the CGIAR or CG system), a

Box 1

Genetically modified technologies in China

China's scientists have generated an impressive array of new technologies. From 353 applications between 1996 and 2000, China's Office of Genetic Engineering Safety Administration approved 251 cases of GM plants, animals and recombinant microorganisms for field trials, environmental releases or commercialization³⁴. Regulators approved 45 GM plant applications for field trials, 65 for environmental release and 31 for commercialization (of which around 20 were various *Bt* cotton varieties). Breakthroughs on food crops that have received little attention elsewhere (over 40% of the trials elsewhere in the world involve GM maize) also demonstrate China's concern for food security as a developing nation (see Table below). Transgenic rice resistant to three of China's main rice pests — stem borer (using *Bt* and cowpea trypsin inhibitor (*CpTI*) genes), planthopper and bacterial leaf blight (using the *Xa21* gene) — have passed at least two years of environmental-release trials. Researchers have moved GM wheat with resistance to barley yellow dwarf virus to field trials, and are experimenting with GM potato and peanut.

Box 1 Table GM plants (commercialized and in trials) in China, 1999

Crop	Introduced trait	Crop	Introduced trait
1. Cotton	Insect resistance*, disease resistance	9. Tobacco	Insect resistance
2. Rice	Insect resistance, disease resistance, herbicide resistance, salt tolerance (BADH)	10. Cabbage	Virus resistance
3. Wheat	BYDV resistance, quality improvement	11. Tomato	Virus resistance*, shelf-life altered*, cold tolerance
4. Maize	Insect resistance (<i>Bt</i>), quality improvement	12. Melon	Virus resistance
5. Soya bean	Herbicide resistance	13. Sweet pepper	Virus resistance*
6. Potato	Disease resistance, quality improvement	14. Chilli	Virus resistance
7. Rape seed	Disease resistance	15. Petunia	Coloured altered*
8. Peanut	Virus resistance	16. Papaya	Virus resistance

*Approved for commercialization; others waiting for commercialization or environmental release.

Abbreviations: BADH, betaine aldehyde dehydrogenase; BYDV, barley yellow dwarf virus. Data from ref. 34.

group of 16 international agricultural research organizations, has taken on the responsibility of transferring technology from industrialized to developing countries and working on crops and research problems that are unique to the tropical and sub-tropical climates. Funded multilaterally, the past successes of the CG system are well documented^{49,50}. In recent years, however, expenditures have fallen in real terms. The system struggles with trying to define a role that fits into the new global economy, meets the needs of developing countries, and is attractive to donor countries in the industrialized world.

Perspectives

Technology generated the production increases in the twentieth century and provided the world with inexpensive food that helped curb malnutrition and alleviate poverty. In the same way that Malthus's observations later in his life moderated his original forecast of global starvation, the record of the Green Revolution certainly has weakened the case put forth by those who had predicted serious food shortages during the postwar era. But despite the increased availability of food, today's world faces food production challenges at least as great as those that faced it several decades ago. Although the magnitude of the problems are hotly debated, many scientists and economists believe the genetic potential of existing varieties are

falling, pests are becoming increasingly difficult to control, and land, water and other resources are becoming scarce.

Fortunately, technological progress, as in the past, has the potential of continuing to aid the world in overcoming many of these problems and meeting its food needs in the twenty-first century. At least for the foreseeable future, scientists in both developing and developed countries have a well-defined vision of the types of technologies that they can create to overcome the constraints that limit greater increases in output. Past partnerships between publicly and privately funded research created the technologies that fuelled the rise in yields and cropping intensity. For future generations of scientist and policy makers to achieve the same success, they must create an environment conducive to creative and rapid technological change.

The creation of a research-friendly environment depends on a number of critical elements. Because research is increasingly expensive and budgets increasingly strained, clear priorities need to be established. The right mix of funding and research effort needs to be apportioned between conventional plant breeding and modern genetic technology. In the short term, conventional plant breeding still has many important roles to play. This is especially true in developing countries, in which the wage levels of scientists are low enough to allow research systems to devote considerable effort to relatively traditional means of experimentation that still hold great potential for creating technological progress. Most developing countries will have difficulty establishing and maintaining strong, modern biotechnology research programmes, as the funding needs and human capital demands often exceed their capabilities. Over the longer run, however, agricultural research will almost inevitably depend on high-technology methods.

Choices also need to be balanced in a number of other ways. There needs to be the right mix of genetics- and agronomic-based research. Scientists need to carefully choose the crops and traits that they work on; the choices will directly affect who benefits and how big of an effect there will be on poverty alleviation. As opportunities arise for private firms to invest in some types of agricultural technologies, the public sector should coordinate its efforts to take advantage of the willingness of the private sector to contribute to the search for new technologies, and focus on those areas that firms are unwilling or unable to invest in.

Ultimately, however, it is going to take a commitment by governments and citizens to take steps to support continued research on agriculture. The funding needs are as great as ever. Support for many types of research have been flagging in recent years. If the struggle to feed billions of new mouths and improve the livelihood of billions of other poor people is to be won, the same commitments that were made in the past to ensure the future of food need to be reaffirmed. □

doi:10.1038/nature01015

- World Bank. *World Development Report, 2000* (World Bank, Washington DC, 2000).
- Sen, A. K. On economics of life and death. *Sci. Am.* **5**, 40–47 (1993).
- Pingali, P. L., Hossain, M. & Gerpacio, R. V. *Asian Rice Bowls* (CAB International, Wallingford, 1997).
- Pachico, D., Hertford, R. & de Janvry, A. Assessing the impact of agricultural research on poverty alleviation. *Food Policy* **25**, 379–388 (2000).
- Pinstrip-Anderson, P., Pandya-Lorch, R. & Rosegrant, M. W. *World Food Prospects: Critical Issues for the Early Twenty-first Century. 2020 Vision Food Policy Report* (International Food Policy Research Institute, Washington DC, 1999).
- Rosegrant, M. W., Leach, N. & Gerpacio, R. V. Alternative futures for world cereal and meat consumption. *Proc. Nutr. Soc.* **58**, 219–234 (1999).
- Byerlee, D., Heisey, P. & Pingali, P. L. in *Food Needs of the Developing World in the Early Twenty First Century* 207–250 (Pontifical Academy of Sciences, Vatican City, 2000).
- Huang, J., Rozelle, S. D. & Rosegrant, M. W. China's food economy to the 21st century: supply, demand, and trade. *Econ. Dev. Cult. Change* **47**, 737–766 (1999).
- Pray, C. E. The green revolution: extensions, institutions, impacts and lessons. Department of Agriculture, Food, and Resource Economics Working Paper (Rutgers University, New Brunswick, NJ, 1998).
- Pingali, P. L. in *Agricultural Technology: Policy Issues for the International Community* (ed. Anderson, J. R.) 384–401 (CAB International, Wallingford, 1994).
- Morris, M. L. & Byerlee, D. in *Agricultural Development: the Third World 3rd edn* (eds Eicher, C. K. & Staatz, J.) 458–473 (Johns Hopkins Press, Baltimore, MD, 1998).
- Cassman, K. G. & Pingali, P. L. in *Agricultural Sustainability: Economic, Environmental and Statistical Considerations* (eds Barnett, V., Payne, R. & Steiner, R.) 63–84 (Wiley, New York, 1995).
- Huang, J. & Rozelle, S. D. Environmental stress and grain yields in China. *Am. J. Agric. Econ.* **77**, 246–256 (1995).
- Cassman, K. G. *et al.* Opportunities for increased nitrogen use for improved resource management in irrigated rice systems. *Field Crop Res.* **56**, 7–39 (1998).
- Ali, M. & Byerlee, D. Productivity growth and resource degradation in Pakistan's Punjab: a decomposition analysis. *Econ. Dev. Cult. Change* (in the press).
- Murgai, R., Ali, M. & Byerlee, D. Productivity growth and sustainability in post-green revolution agriculture: the case for the Indian and Pakistan Punjab. *World Bank Res. Observ.* **16**, 199–218 (2001).
- World Bank. *World Development Report, 1992* (World Bank, Washington DC, 1992).
- Barker, R., Herdt, R. W. & Rose, B. *Rice Economy of Asia* (Resources For the Future, Washington DC, and International Rice Research Institute, Los Baños, Laguna, Philippines, 1985).
- Evenson, R. E., Herdt, R. W. & Hossain, M. *Rice Research in Asia: Progress and Priorities* (CAB International, Wallingford, 1996).
- Pingali, P. L. & Rola, A. C. *Public Regulatory Roles in Developing Markets: The Case of Pesticides* (International Rice Research Institute, Los Baños, Laguna, Philippines, 1994).
- Antle, J. M. & Pingali, P. L. Pesticides, productivity and farmer health: a Philippine case study. *Am. J. Agric. Econ.* **76**, 418–430 (1994).
- Fan, S., Hazell, P. B. R. & Thorat, S. *Linkages between Government Spending, Growth and Poverty in Rural India* Res. Rep. No. 110 (International Food Policy Research Institute, Washington DC, 1999).
- Fan, S., Zhang, L. & Zhang, X. *Growth, Inequality, and Poverty in Rural China: The Role of Public Investments* Res. Rep. No. 125 (International Food Policy Research Institute, Washington DC, 2002).
- Huang, Q., Rozelle, S. D., Huang, J. & Wang, J. Irrigation and yields in China's agriculture. Department of Agricultural and Resource Economics Working Paper (Univ. California, Davis, 2002).
- Seckler, D., Molden, D. & Barker, R. *Water Scarcity in the 21st Century* (International Water Management Institute, Colombo, 1998).
- Virmani, S. S., Khush, G. S. & Pingali, P. L. in *Hybrid Research and Development Needs in Major Cereals in the Asia-Pacific Region* (eds Paroda, R. S. & Rai, M.) 61–86 (Food and Agriculture Organization of the United Nations, Regional Office for Asia and the Pacific, Bangkok, 1994).
- Huang, J. & Rozelle, S. D. Technological change: rediscovering of the engine of productivity growth in China's rural economy. *J. Dev. Econ.* **49**, 337–369 (1996).
- Yuan, L. & Lu, X. Two-line rice hybridization technology and intermedium test development program. *Biol. Eng. Prog.* **20**, 2–6 (2000). [In Chinese.]
- Khush, G. S. Modern varieties: their real contribution to food supply and equity. *Geojournal* **35**, 275–284 (1995).
- David, C. C. & Otsuka, K. *Modern Rice Technology and Income Distribution in Asia*. (International Rice Research Institute, Los Baños, Laguna, Philippines, and Lynne Rienner Publishers, Boulder and London, 1994).
- James, C. A. *Global Status of Commercialized Transgenic Crops, 1999* ISAAA Briefs No. 17–2000 (International Service for the Acquisition of Agri-biotech Applications, Ithaca, NY, 2000).
- James, C. A. *Global Review of Commercialized Transgenic Crops, 2001* ISAAA Briefs No. 23–2001 (International Service for the Acquisition of Agri-biotech Applications, Ithaca, NY, 2001).
- Qaim, M. & Traxler, G. S. Roundup-Ready soybeans in Argentina: farm level, environmental, and welfare effects. International Conference on Agricultural Biotechnology Research, Ravenna, Italy, 11–14 July 2002 <<http://www.economia.uniroma2.it/conferenze/icabr/abstract/Qaim.htm>> (2002).
- Huang, J., Rozelle, S. D., Pray, C. E. & Wang, Q. Plant biotechnology in China. *Science* **295**, 674–677 (2002).
- Traxler, G., Godoy-Avila, S., Falck-Zepeda, J. & Espinoza-Arellano, J. Transgenic cotton in Mexico: economic and environmental impacts. Department of Agricultural Economics Working Paper (Auburn Univ., Auburn, AL, 2001).
- Ismael, Y. Smallholder adoption and economic impacts of Bt cotton in the Makhathini Flats, Republic of South Africa. Report for DFID Natural Resources Policy Research Programme, Project R7946 (Department for International Development, London, 2001).
- Huang, J., Wang, Q. & Zhang, Y. Agricultural biotechnology development and research capacity. Center for Chinese Agricultural Policy Working Paper (Center for Chinese Agricultural Policy, Chinese Academy of Sciences, Beijing, 2001).
- Pray, C. E. & Schimmpfennig, D. E. Role of biotech in South Africa's agriculture. Department of Agriculture, Food and Resource Economics Working Paper (Rutgers University, New Brunswick, NJ, 2001).
- Pray, C. E., Courtmanche, A. & Govindasamy, R. The importance of intellectual property rights in the international spread of private sector agricultural biotechnology. International Conference on Agricultural Biotechnology Research, Ravenna, Italy, 11–14 July 2002 <<http://www.economia.uniroma2.it/conferenze/icabr/abstract/Pray1.htm>> (2002).
- Ekboir, J. (ed.) *World Wheat Overview and Outlook 2000–2001: Developing No-Till Packages for Small-Scale Farmers* (International Maize and Wheat Improvement Center (CIMMYT), Mexico, 2001).
- Khanna, M., Epouhe, O. F. & Hornbaker, R. Site-specific crop management: adoption patterns and incentives. *Rev. Agric. Econ.* **21**, 455–472 (1999).
- Alston, J. M., Pardey, P. G. & Roseboom, J. Financing agricultural research: international investment patterns and policy perspectives. *World Dev.* **26**, 1057–1072 (1998).
- Pardey, P. G. & Beintema, N. M. *Slow Magic: Agricultural R&D A Century after Mendel* (International Food Policy Research Institute, Washington DC, 2001).
- Alston, J. M., Marra, M. C., Pardey, P. G. & Wyatt, T. J. Research returns redux: a meta-analysis of rates of return to agricultural R&D. *Aust. J. Agric. Resource Econ.* **44**, 185–215 (2000).
- Huffman, W. E. & Evenson, R. E. *Science for Agriculture: A Long-Term Perspective* (Iowa State Univ. Press, Ames, 1993).
- Pray, C. E. Public/private sector linkages in research and development: biotechnology and the seed industry in Brazil, China and India. *Am. J. Agric. Econ.* **83**, 742–747 (2001).
- Pray, C. E. & Umali-Deininger, D. Private agricultural research: Will it fill the gap? *World Dev.* **26**, 1127–1148 (1998).
- Jin, S., Huang, J., Hu, R. & Rozelle, S. D. The creation and spread of technology and total factor productivity in China. *Am. J. Agric. Econ.* (in the press).
- Evenson, R. E. & Gollin, D. A Comprehensive assessment of the role of the CG system in increasing yields in developing countries. Economic Growth Center Working Paper (Yale Univ. Growth Center, New Haven, CT, 2001).
- Serageldin, I. & Persley, G. J. *Promethian Science: Agricultural Biotechnology, the Environment, and the Poor* (Consultative Group on International Agricultural Research, Washington DC, 2000).

Assessing the risks associated with new agricultural practices

R. S. Hails

NERC Centre for Ecology and Hydrology, Mansfield Road, Oxford OX1 3SR, UK (e-mail: rha@ceh.ac.uk)

One key challenge for the twenty-first century is how to produce the food we need, yet ensure the landscape we want. Genetically modified crops have focused our attention on how to answer this question for one part of agriculture. The same principles could be applied to assess environmental impacts of future land-use change in a much broader context.

The improvement of crop varieties through plant breeding has greatly increased yields over the past few decades. This in turn has reduced the area of land that would otherwise have been dedicated to agriculture as a result of an increasing world population¹. But these new crop varieties have been accompanied by changes in agronomic practice, which have included increased use of fertilizers, pesticides and irrigation. Many of these changes have been to the detriment of wildlife². Change has therefore brought both benefits and costs, and the future of food is linked inextricably with the future of the environment.

In this context, biotechnology offers new possibilities for the twenty-first century. Will the introduction of genetically modified (GM) crops exacerbate current trends or offer new avenues for environmental management? The debate presents regulators and scientists with a significant challenge — how to define and assess impact against a backdrop of constantly shifting land-management practices and resultant biodiversity. The approaches being developed in response to this debate are of generic value. Ecological science has much to offer in providing objective evidence of potential impacts of land-management practices. With sufficient resources, future changes in land management could be directed towards specific environmental goals. Solutions, however, will require an integration of science, policy and regulation.

One important issue is the proportion of wildlife that is dependent upon the management of agricultural land. In the United States and Canada, large areas are dominated by farming, but large areas are also set aside for recreation and conservation. In contrast, in Europe, agriculture and the countryside are intimately intertwined, with around 70% of land area classified as agricultural land (Fig. 1). Europeans effectively live inside their national parks. This may explain some differences in perspective towards the role of agriculture in delivering wildlife, and why the potential introduction of GM plants has raised some complex issues in Europe.

Risk assessment of GM plants has been divided traditionally into direct and indirect impacts. Direct impacts arise from the presence of the transgenic plant itself, or the consequences of transfer of the transgene into wild relatives. Indirect impacts arise from the management practices associated with the transgenic crop. Ecological theory has an important role to play in assessing such impacts.

Assessing the direct impact of GM crops

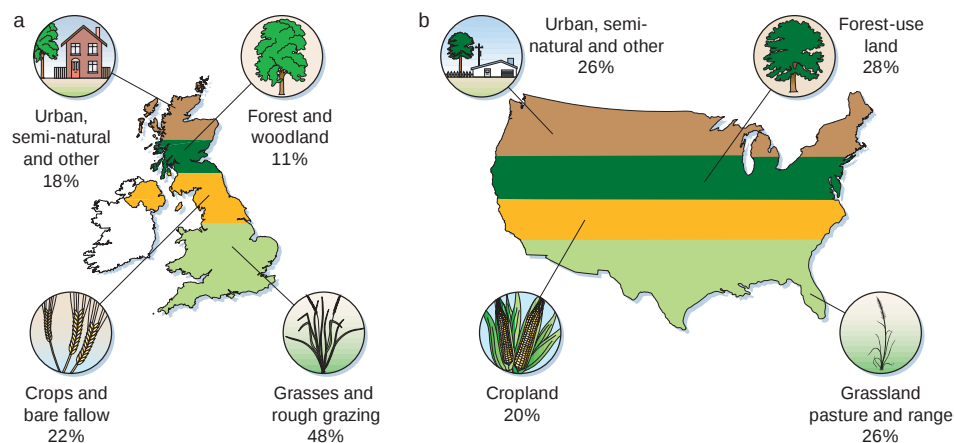
The potential direct impacts of a GM plant include any changes in ecological fitness which may make the crop plant

or any crop plant/wild relative hybrids more invasive (see ref. 3 for a definition of ecological fitness). Invasive species represent one of the greatest threats to biodiversity^{4–6}. Most non-native species are harmless, but a few become invasive and detrimental to indigenous ecosystems. Japanese knotweed (*Fallopia*) and rhododendron (*Rhododendron ponticum*) are notable examples in the United Kingdom, while yellow star thistle (*Centaurea solstitialis*) and European purple loosestrife (*Lythrum salicaria*) are particularly disruptive examples in the United States. Most of these invasive species originate from horticulture, and so far there is no evidence that any GM crop plant is significantly more invasive than its conventional counterpart⁷.

Nevertheless, the possibility exists that certain transgenic plants could pose a direct threat through enhanced ecological fitness. In particular, transgenes that confer enduring resistance to pathogens could cause ecological release in wild relatives if populations are regulated by those same pathogens. Indeed, the ethos behind biological control relies on this principle, in that it assumes that some plants become weeds because they have escaped their natural enemies, and seeks to reduce populations by the introduction of a suitable pathogen or herbivore. Wild plant populations contain a variety of resistance genes that have evolved in response to the presence of pathogens. But genetic modification greatly widens the pool of potential resistance genes, allowing the use of new pathogen-resistance mechanisms. The use of viral coat protein genes to provide resistance against the same viral strains is a case in point⁸. Such resistance mechanisms may prove to be more enduring than those produced by conventional plant breeding. Although this is good news for agriculture, the potential for causing ecological release in wild populations should be considered.

Attempting to predict which plants will prove to be invasive in advance of their release still involves a great deal of uncertainty, and there is nothing to compare with addressing this issue through focussed field experiments, comparing ecological fitness between transgenics and their conventional counterparts^{7,9,10}. The design can involve manipulation to greatly widen the range of ecological conditions, and should be conducted over several years to encompass a range of abiotic conditions. This direct approach can draw on ecological theory to focus on those genotypes most likely to invade, and on those habitats that are most likely to be vulnerable to invasion¹¹. This should provide a good indication of potentially invasive plants, but as with all such predictions, there will always remain a

Figure 1 Land use by agriculture and other uses. **a**, Land use for the United Kingdom in 2000. The total area classified as agricultural land is 70%, including both cropland and grassland/rough grazing. Source: ref. 50. **b**, Principal uses of land in the United States in 1997. Source: ref. 51.



degree of uncertainty. The uncertainty we are prepared to accept in the introduction of GM crop plants should not be out of all proportion to that we accept for conventional crop plants, let alone imported exotics⁵. The approaches developed to answer such questions for GM plants could be profitably applied to imported plant species.

Assessing the indirect impacts of GM crops

The much more challenging question is how do we assess the potential impacts of changing agricultural practice? This article will focus on GMHT (genetically modified, herbicide tolerant) crops as a case study, for which any impacts are most likely to be through changes in the efficiency of weed management.

GMHT crops are resistant to broad-spectrum herbicides such as glyphosate or glufosinate. Consequently, these herbicides may be applied after crop germination without harming the GM crop itself. But because these herbicides have a broad range of action, their use may reduce weed and invertebrate populations, on which birds and other wildlife depend. If herbicide application is instead delayed until later in the growing season, just before the point where weed populations would start to compete with the crop, a greater diversity and biomass of in-field vegetation would occur early in the season, during nesting time for many birds. However, fewer plants may survive to set seed after herbicide application, so this may reduce the abundance and diversity of seedlings the following year.

ACRE, the UK government's Advisory Committee on Releases to the Environment, has developed guidelines on how to assess the indirect impacts of a GM crop¹². They require a comparison between the management of the GM crop and the equivalent non-GM crop, with an assessment of the potential impact on key indicator species typical of arable farmland. Such comparisons can be meaningfully made only through large-scale field experiments, and in fact these guidelines build on the experience of the farm-scale evaluations (FSEs), set up in the United Kingdom in 1999. The FSEs are designed to investigate the potential indirect impacts of four GMHT crops (spring and winter oilseed rape, maize and sugar beet). They are funded by the UK government, executed by a consortium of research institutes, and overseen by an independent scientific steering committee¹³. The design of these trials is outlined in Box 1, and the first set of results will be published in 2003.

These ecological experiments have provoked considerable interest from the public, and initially were presented as providing the final word in the risk assessment of these particular crops¹⁴. It is now more widely recognized that they will inform the debate, but that our options are considerably more complex than a simple pass or fail for this technology¹⁴. The trials will provide valuable data on the impact of the management of both GM and conventional crops

on key indicator species, and large-scale manipulative experiments represent the most direct and reliable method of obtaining this information. In this context, the trials are pioneering in their attempt to link a proposed change in land management with potential changes in biodiversity (number of species), prior to its introduction.

Such trials should become a cornerstone of our strategy if we aim to use ecological science to inform future agricultural and environmental strategy. But as with many good experiments, they will raise as many issues as they solve. The interpretation of the data is unlikely to be clear and straightforward — biodiversity indicators may be enhanced during certain parts of the year, and reduced at other times. Some species will become more abundant, while others will be reduced. Do we value some species over others (skylarks over earthworms for example)? And how do we extrapolate from changes in abundance of weeds and invertebrates at different times of year to potential changes in bird and mammal populations?

Mathematical models provide one means of scaling up our predictions, both spatially and temporally. Watkinson *et al.*¹⁵ provide an elegant example of this, with a model that extrapolates from data on weed numbers in conventional and GMHT crops, to calculate the impact of herbicide use on farmland bird numbers. Such models may be used to investigate the potential impacts of a range of management scenarios, and could aid the design of management guidelines¹⁶. Data could then be gathered in such a way that could test any predictions, in the early stages of commercialization. The long-term monitoring that would be required post-commercialization under European Union legislation could be designed for precisely this purpose.

Developing protocols for risk assessment

There has been some debate in the literature about how to manage the world's ecosystems as capital assets, which is highly pertinent to the debate on the future of agriculture^{17,18}. In looking forward to the coming decades, we can continue to develop the more proactive approach of the FSEs to consider land-use change more broadly. The first step would be to specify our goals. For example, do we intend to farm for biodiversity as well as food? This requires decisions to be made about which species or habitats we wish to preserve within the agro-ecosystem (as opposed to nature reserves)¹⁹, and the extent to which we are prepared to pay for the 'non-food' component of agriculture²⁰. Clearly these biodiversity goals need to be integrated with the goal of affordable, safely produced, high-quality food.

Perhaps most important, agro-ecosystem biodiversity is only one piece of the jigsaw. In organic and many conventional systems, weeds are controlled by ploughing before planting the crop. GMHT crops could support a reduction in tillage through direct drilling into a weedy field, which may be beneficial to soil organisms. The glyphosate

Box 1

The farm-scale evaluations

The farm-scale evaluations are a series of experiments conducted over three years to determine the potential impact of GMHT (genetically modified, herbicide tolerant) crops on the agro-ecosystem.

The null hypothesis

There are no statistically significant differences in the abundance and diversity of farmland wildlife associated with the management of GMHT crops compared with the management of equivalent non-GM crops.

A number of sensitive indicator species have been chosen to represent 'wildlife' from the following groups: all plants, soil surface active arthropods, plant arthropods and gastropods, bees, butterflies and the soil seed bank. The emphasis is very much on species at the lower end of the food chain, as treatments applied at the field scale are unlikely to have much impact on mobile species such as birds and mammals, and with the anticipation that knock-on effects to larger organisms that rely on these species for food can be deduced through models^{15,16}.

The design

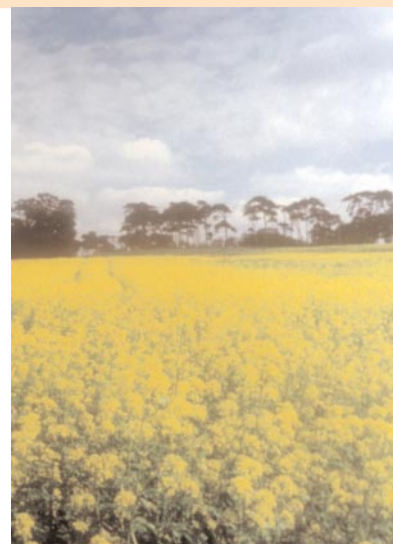
Preliminary data suggested that a split-field design would be more powerful and

appropriate than a paired-field design. Consequently, each experimental field (ranging in size from 3–12 hectares) is divided into two, and the GMHT crop is grown in one half (chosen at random), with an equivalent conventional crop grown in the other half. This minimizes any differences in biodiversity at the start of the experiment.

The management of the GM crop follows guidelines laid down by SCIMAC (Supply Chain Initiative on Modified Arable Crops), an industry body that advises on cost-effective weed control in the GM crop (subsequently audited by the scientists).

Sampling takes place throughout the year in the crop, headland and field margins. Power analysis is used to determine the size of the experiment (the number of farms needed to participate), to ensure with high probability that a 50% change in abundance of any one of the common species would be detected.

By the close of the experiment, around 60 farms will have taken part across the United Kingdom for each of the four crops, chosen to represent the range of regional geographical differences and current farming methods.



The funding and management

These trials are funded by the UK government, undertaken by a consortium of research institutes, and overseen by an independent scientific steering committee. The results will be available in 2003. For more information see:

♦ www.defra.gov.uk/environment/fse/index.html/.

or glufosinate herbicides used with these crops are less persistent than many conventional herbicides such as atrazine, and may lead to a reduction in the costs of removing agrochemicals from watercourses. Moving from herbicide-tolerant to insect-protected GM varieties, the potential to reduce pesticide use is even greater²¹. Fewer herbicide or pesticide treatments will also mean reduced consumption of fossil fuels driving agricultural machinery and a significant reduction in atmospheric carbon dioxide emissions²¹. A true comparison of environmental impact would require all externalities to be included²², encompassing ecological, social and economic considerations.

The second step is to evaluate all possible routes to achieve our goals. This should include all ways in which GM crops could be managed to minimize environmental impact, as well as all alternative agricultural systems. In the context of GM crops, the FSEs are designed to answer the question 'What will be the potential impact of the management of GMHT crops on key biodiversity indicators, compared with the equivalent conventional herbicide-tolerant crop?' This question informs us about impact in the context of current agricultural practice. However, an equally valid question might be 'In what ways can GMHT crops be managed to maximize biodiversity without reducing yield?' Other field trials are currently addressing some of these issues in the United Kingdom²³, and illustrate the importance of the framing of the question.

It would be illogical to focus on one particular technology, such as GM crops, if our aim is to deliver wildlife as well as food. The potential exists to reduce pesticide use in a variety of ways, including enhancing beneficial insects and birds, inducing the plant's own innate defences against pests or diseases through chemical signals, using pheromone traps to interfere with mate-finding, and resorting to chemicals only when specific thresholds are reached. The philosophy of integrated pest management has been evident for the past few decades, and is

promoted through a variety of organizations (for example, the UK charity Linking Environment And Farming or LEAF²⁴), but needs to become mainstream policy to maximize its contribution to sustainable change. Technological advances such as precision agriculture, in which fertilizer applications rates are adjusted across fields according to crop needs, could contribute to this process²⁵.

The third step would be an evaluation of the status quo. One signal is clear — biodiversity is in decline. This is a global trend, and many of these changes may be ascribed to agriculture (for example, changes in the floral diversity of grasslands and arable field margins resulting from overgrazing or fertilizer input^{26,27}, loss of critical habitats through eutrophication and conversion to agriculture²⁸, and loss of species diversity within the agro-ecosystem through changes in management²⁹). Restoration ecology is the science of restoring species diversity through focused management strategies, and considerable success has been achieved in the restoration of some habitats, such as species-rich grasslands. Yet not all such losses may be restored. Abiotic constraints (for example, the degree of eutrophication and acidification) and biotic constraints (for example, an impoverished seed bank) may make restoration unfeasible²⁷.

Farmland birds rely on seeds and invertebrates, and so are considered good indicators of farmland biodiversity more widely¹⁵, but there are practical difficulties in studying biodiversity changes for such species at higher trophic levels. Being highly mobile, they respond to land-use changes at much greater spatial scales. Nevertheless, taking Europe as an example, correlative studies illustrate that conservationally valued farmland birds are in decline, and this seems to be associated with an intensification of agriculture in the latter part of the 20th century^{30–33}. One hypothesis is that declines may be associated with changes in frequency of those arable plants that are important food resources for farmland birds^{34,35}. Countryside

surveys, conducted across the United Kingdom over the past 25 years support this hypothesis, concluding that arable field centres are a much poorer source of food plants for farmland birds now than they were in 1978^{36–38}.

Such simple temporal correlations may be misleading, and more detailed knowledge of the life history of a bird species is required to link cause and effect. Ecological studies have provided this in some cases. The switch from growing spring- to autumn-sown crops deprives many species of feeding sites in stubble fields overwinter^{39–41}. Grass margins around arable fields and weedy winter stubbles, enhanced with countryside stewardship schemes, have greatly benefited the threatened curlew⁴². Skylarks breed more successfully in spring cereals and set-aside than in winter cereals, as a result of the impact of vegetation height during the latter part of the breeding season⁴³.

There is clearly an ecological cost to remaining with the status quo, and there are considerable pressures for changes in agricultural policy to address these costs⁴⁴. The fourth and final step will be to direct changes in land use towards our environmental goals. This will require both financial incentives, and up-to-date information on the best options available. Both government and environmental organizations favour shifts in subsidies towards environmental targets⁴⁵. However, resources are needed for research to find the optimal solutions that address the local conditions, and for government-sponsored extension agents to communicate the best available options to the farmer as they come on-line.

Conclusions

The challenge is to produce the food we need for the future while minimizing the environmental impact. This future vision of agriculture involves the farmer as a steward of the countryside⁴⁶. If we put aside the ideological boundaries that are often set up between different agricultural systems (organic, conventional and GM), the objectivity of ecological science can be used to pick the best elements of all systems. It is equally vital that stewardship schemes are designed with scientific input if goals are to be reached^{19,47}. Currently about €1.7 billion (\$US1.7 billion) are spent annually in Europe on agri-environment schemes, yet there are very little data on the effectiveness of these programmes⁴⁸. Long-term monitoring of key elements of biodiversity should be carefully designed, with the validation of these schemes in mind. Without long-term data we cannot direct subsidies to the most effective schemes, detect trends, or assess future changes in biodiversity.

We should promote an experimental approach to elucidate the consequences of specific changes in land management. The FSEs provide a good example of a 'proactive experiment', a product of science-based risk-assessment protocols, the aim of which is to ensure that any future change is in a better direction. This will be most effectively developed in the future if we throw off the constraints of our focus on GM technology, and use the approach to address other land-management issues. Such experiments need to be designed and resourced at the appropriate scale, so that we can adequately address these major practical questions⁴⁹. □

doi:10.1038/nature01016

- Conway, G. & Toennissen, G. Feeding the world in the twenty-first century. *Nature* **402**, C55–C58 (1999).
- Krebs, J. R., Wilson, J. D., Bradbury, R. B. & Sirdwardena, G. M. The second Silent Spring? *Nature* **400**, 611–612 (1999).
- Crawley, M. J., Hails, R. S., Rees, M., Kohn, D. & Buxton, J. The ecology of transgenic oilseed rape in natural habitats. *Nature* **363**, 620–623 (1993).
- Sala, O. E. et al. Global biodiversity scenarios for the Year 2100. *Science* **287**, 1770–1774 (2000).
- Trewavas, A. J. & Leaver, C. J. Is opposition to GM crops science or politics? *EMBO Rep.* **2**, 1–5 (2001).
- Pimentel, D., Lach, L., Zuniga, R. & Morrison, D. Environmental and economic costs of nonindigenous species in the United States. *Bioscience* **50**, 53 (2000).
- Crawley, M. J., Brown, S. L., Hails, R. S., Kohn, D. & Rees, M. The performance of transgenic crops in natural habitats: a 10-year perspective. *Nature* **409**, 682–683 (2001).
- Cooper, J. I. in *Commercialisation of Transgenic Crops: Risk, Benefit and Trade Considerations* (eds McLean, G. D. et al.) 115–124 (Bureau of Resource Sciences, Kingston, ACT, 1997).
- Crawley, M. J. Bollworms, genes and ecologists. *Nature* **400**, 501–502 (1999).

- Hails, R. S. Genetically modified organisms—the debate continues. *Trends Ecol. Evol.* **15**, 14–18 (2000).
- Shea, K. & Chesson, P. What can community theory tell us about biological invasions? *Trends Ecol. Evol.* **17**, 170–176 (2002).
- Advisory Committee on Releases to the Environment (ACRE) Sub-group on Wider Biodiversity Issues. *Guidance on the Assessment of the Impact on Wider Biodiversity from Proposed Cultivation of GM Crops* <http://www.defra.gov.uk/environment/acre/biodiversity/guidance/index.htm> (31 July 2001).
- Firbank, L. G. et al. Farm-scale evaluation of GM crops explained. *Nature* **399**, 727–728 (1999).
- Agriculture and Environment Biotechnology Commission. *Crops on Trial* (DTI, London, 2001).
- Watkinson, A. R., Freckleton, R. P., Robinson, R. A. & Sutherland, W. J. Predicting biodiversity responses to GM herbicide tolerant crops. *Science* **289**, 1554–1556 (2000).
- Firbank, L. G. & Forcella, F. Genetically modified crops and farmland biodiversity. *Science* **289**, 1481–1482 (2000).
- Daily, G. C. et al. The value of Nature and the nature of value. *Science* **289**, 395–396 (2000).
- Tilman, D. et al. Forecasting agriculturally driven global environmental change. *Science* **292**, 281–284 (2001).
- Sutherland, W. J. Restoring a sustainable countryside. *Trends Ecol. Evol.* **17**, 148–150 (2002).
- Beringer, J. E. Farming for biodiversity with GMOs? *Outlook Agric.* **30**, 7–9 (2001).
- Phipps, R. H. & Park, J. R. Environmental benefits of genetically modified crops: global and European perspectives on their ability to reduce pesticide use. *J. Anim. Feed Sci.* **11**, 1–18 (2002).
- Pretty, J. N. et al. An assessment of the total external costs of UK agriculture. *Agric. Syst.* **65**, 113–136 (2000).
- Pidgeon, J. D., Dewar, A. M. & May, M. J. in *Proc. BCPC-Weeds 2001*, 373–380 (BCPC Publications, Bracknell, 2001).
- Marsh, J. *Integrated Farm Management—A Farm Strategy for the 21st Century* (LEAF, The National Agricultural Centre, Stoneleigh, 2000).
- Tilman, D. The greening of the green revolution. *Nature* **396**, 211–212 (1998).
- Wilson, P. J. in *Field Margins: Integrating Agriculture and Conservation* (ed. Boatman, N.) BCPC Monogr. No. 58, 253–258 (BCPC Publications, Bracknell, 1994).
- Bakker, J. P. & Berendse, F. Constraints in the restoration of ecological diversity in grassland and heathland communities. *Trends Ecol. Evol.* **14**, 63–68 (1999).
- Bobbink, R., Hornung, M. & Roelofs, J. G. M. in *Manual on Methodologies and Criteria for Mapping Critical Loads/Levels* (UN ECE Convention on Long-range Transboundary Air Pollution), 71–96, III-i-III-54 (Federal Environmental Agency, Berlin, 1996).
- Perfecto, I., Rice, R. A., Greenberg, R. & Van der Voort, M. E. Shade coffee: a disappearing refuge for biodiversity. *Bioscience* **46**, 598–608 (1996).
- Chamberlain, D. E., Fuller, R. J., Bunce, J. C., Duckworth, J. C. & Shrubbs, M. Changes in the abundance of farmland birds in relation to the timing of agricultural intensification in England and Wales. *J. Appl. Ecol.* **37**, 771–788 (2000).
- Chamberlain, D. E. & Fuller, R. J. Local extinctions and changes in species richness of lowland farmland birds in England and Wales in relation to recent changes in agricultural land-use. *Agric. Ecosyst. Environ.* **78**, 1–17 (2000).
- Gates, S. & Donald, P. F. Local extinction of British farmland birds and the prediction of further loss. *J. Appl. Ecol.* **37**, 806–820 (2000).
- Donald, P. F., Green, R. E. & Heath, M. F. Agricultural intensification and the collapse of Europe's farmland bird populations. *Proc. R. Soc. Lond. B* **268**, 25–29 (2001).
- Campbell, L. H. et al. *A Review of the Indirect Effect of Pesticides on Birds* INCC Report No. 227 (Joint Nature Conservation Committee, Peterborough, 1997).
- Wilson, J. D., Morris, A. J., Arroyo, B. E., Clark, S. C. & Bradbury, R. B. A review of the abundance and diversity of invertebrate and plant foods of granivorous birds in northern Europe in relation to agricultural change. *Agric. Ecosyst. Environ.* **75**, 13–30 (1999).
- Haines-Young, R. H. et al. *Accounting for Nature: Assessing Habitats in the UK Countryside* (Department of the Environment, Transport and the Regions, London, 2000).
- Smart, S. M., Firbank, L. G., Bunce, R. G. H. & Watkins, J. W. Quantifying changes in abundance of food plants for butterfly larvae and farmland birds. *J. Appl. Ecol.* **37**, 398–414 (2000).
- Firbank, L. & Smart, S. in *Aspects of Applied Biology, Birds and Agriculture* (eds Boatman, N. D. et al.) 165–170 (Association of Applied Biologists, Wellesbourne, 2002).
- Hald, A. B. The impact of changing the season in which cereals are sown on the diversity of the weed flora in rotational fields in Denmark. *J. Appl. Ecol.* **36**, 24–32 (1999).
- Sirdwardena, G. M., Baille, S. R. & Wilson, J. D. Variation in the survival rates of some British passerines with respect to their population trends on farmland. *Bird Study* **45**, 276–292 (1998).
- Sirdwardena, G. M., Baille, S. R. & Wilson, J. D. Temporal variation in the annual survival rates of six granivorous birds with contrasting population trends. *Ibis* **141**, 621–636 (1999).
- Peach, W. J., Lovett, L. J., Wotton, S. R. & Jeffs, C. Countryside stewardship delivers curlew buntings (*Emberiza cirius*) in Devon, UK. *Biol. Conserv.* **101**, 361–373 (2001).
- Chamberlain, D. E., Wilson, A. M., Browne, S. J. & Vickery, J. A. Effects of habitat type and management on the abundance of skylarks in the breeding season. *J. Appl. Ecol.* **36**, 856–870 (1999).
- Robinson, R. A. & Sutherland, W. J. Post-war changes in arable farming and biodiversity in Great Britain. *J. Appl. Ecol.* **39**, 157–176 (2002).
- Falconer, K. & Ward, N. Using modulation to green the CAP: the UK case. *Land Use Policy* **17**, 269–277 (2000).
- Policy Commission on the Future of Farming and Food. *Farming and Food—A Sustainable Future* <http://www.cabinet-office.gov.uk/farming/index/CommissionReport.htm> (January 2002).
- Kleijn, D., Berendse, F., Smit, R. & Gilissen, N. Agri-environment schemes do not effectively protect biodiversity in Dutch agricultural landscapes. *Nature* **413**, 723–725 (2001).
- Whitfield, J. Birds fly in the face of 'green' farming incentive scheme. *Nature* **413**, 659 (2001).
- Lawton, J. H. *Community Ecology in a Changing World* Excellence in Ecology no. 11 (International Ecology Institute, Oldendorf/Luhe, 2000).
- Department for Environment, Food and Rural Affairs. *Digest of Environmental Statistics* <www.defra.gov.uk/environment/statistics/des/index.htm> (November 2001).
- Vesterby, M. & Krupa, K. S. *Major Uses of Land in the US, 1997* ERS Statistical Bulletin No. 973 <http://www.ers.usda.gov/publications/sb973/> (September 2001).

Acknowledgements

Many thanks to J. Perry, A. Gray, A. Lilley, P. Dale and L. Firbank for helpful comments and discussion.

Towards sustainability in world fisheries

Daniel Pauly, Villy Christensen, Sylvie Gu  nette, Tony J. Pitcher, U. Rashid Sumaila, Carl J. Walters, R. Watson & Dirk Zeller

Fisheries Centre, University of British Columbia, 2204 Main Mall, Vancouver, British Columbia, Canada V6T 1Z4
(e-mail: d.pauly@fisheries.ubc.ca)

Fisheries have rarely been 'sustainable'. Rather, fishing has induced serial depletions, long masked by improved technology, geographic expansion and exploitation of previously spurned species lower in the food web. With global catches declining since the late 1980s, continuation of present trends will lead to supply shortfall, for which aquaculture cannot be expected to compensate, and may well exacerbate. Reducing fishing capacity to appropriate levels will require strong reductions of subsidies. Zoning the oceans into unfished marine reserves and areas with limited levels of fishing effort would allow sustainable fisheries, based on resources embedded in functional, diverse ecosystems.

Fishing is the catching of aquatic wildlife, the equivalent of hunting bison, deer and rabbits on land. Thus, it is not surprising that industrial-scale fishing should generally not be sustainable: industrial-scale hunting, on land, would not be, either. What is surprising rather, is how entrenched the notion is that unspecified 'environmental change' caused, and continues to cause, the collapse of exploited fish populations. Examining the history of fishing and fisheries makes it abundantly clear that humans have had for thousands of years a major impact on target species and their supporting ecosystems¹. Indeed, the archaeological literature contains many examples of ancient human fishing associated with gradual shifts, through time, to smaller sizes and the serial depletion of species that we now recognize as the symptoms of overfishing^{1,2}.

This literature supports the claim that, historically, fisheries have tended to be non-sustainable, although not unexpectedly there is a debate about the cause for this³, and the exceptions⁴. The few uncontested historical examples of sustainable fisheries seem to occur where a superabundance of fish supported small human populations in challenging climates⁵. Sustainability occurred where fish populations were naturally protected by having a large part of their distribution outside of the range of fishing operations. Hence, many large old fecund females, which contribute overwhelmingly to the egg production that renews fish populations, remained untouched. How important such females can be is illustrated by the example of a single ripe female red snapper, *Lutjanus campechanus*, of 61 cm and 12.5 kg, which contains the same number of eggs (9,300,000) as 212 females of 42 cm and 1.1 kg each⁶. Where such natural protection was absent, that is, where the entire population was accessible to fishing gears, depletion ensued, even if the gear used seems inefficient in retrospect^{7,8}. This was usually masked, however, by the availability of other species to target, leading to early instances of depletions observable in the changing size and species composition of fish remains, for example, in middens⁹.

The fishing process became industrialized in the early nineteenth century when English fishers started operating steam trawlers, soon rendered more effective by power winches and, after the First World War, diesel engines¹⁰. The

aftermath of the Second World War added another 'peace dividend' to the industrialization of fishing: freezer trawlers, radar and acoustic fish finders. The fleets of the Northern Hemisphere were ready to take on the world.

Fisheries science advanced over this time as well: the two world wars had shown that strongly exploited fish populations, such as those of the North Sea, would recover most, if not all, of their previous abundance when released from fishing¹¹. This allowed the construction of models of single-species fish populations whose size is affected only by fishing pressure, expressed either as a fishing mortality rate (F , or catch/biomass ratio), or by a measure of fishing effort (f , for example, trawling hours per year) related to F through a catchability coefficient^{12,13} (q): $F = qf$. Here, q represents the fraction of a population caught by one unit of effort, directly expressing the effectiveness of a gear. Thus, q should be monitored as closely as fishing effort itself, if the impact of fishing on a given stock, as expressed by F , is to be evaluated. Technology changes tend to increase q , leading to increases referred to as 'technology coefficient'¹⁴, which quickly renders meaningless any attempts to limit fishing mortality by limiting only fishing effort.

The conclusion of these models, still in use even now (although in greatly modified forms; Box 1), is that adjusting fishing effort to some optimum level should generate 'maximum sustainable' yield, a notion that the fishing industry and the regulatory agencies eagerly adopted — if only in theory¹⁵. In practice, optimum effort levels were very rarely implemented (the Pacific halibut fishery is one exception¹⁶). Rather the fisheries expanded their reach, both offshore, by fishing deeper waters and remote sea mounts¹⁷, and by moving onto the then untapped resources of West Africa¹⁸, southeast Asia¹⁹, and other low-latitude and Southern Hemispheric regions²⁰.

Fisheries go global

In 1950, the newly founded Food and Agriculture Organization (FAO) of the United Nations began collection of global statistics. Fisheries in the early 1950s were at the onset of a period of extremely rapid growth, both in the Northern Hemisphere and along the coast of the countries of what is now known as the developing world. Everywhere that industrial-scale fishing (mainly trawling, but also purse seining

and long-lining) was introduced, it competed with small-scale, or artisanal fisheries. This is especially true for tropical shallow waters (10–100 m), where artisanal fisheries targeting food fish for local consumption, and trawlers targeting shrimps for export, and discarding the associated by-catch, compete for the same resource²¹.

Box 1

Single-species stock assessments

Single-species assessments have been performed since the early 1950s, when the founders of modern fisheries science^{12,13} attempted to equate the concept of sustainability with the notion of optimum fishing mortality, leading to some form of maximum sustainable yield. Most of these models, now much evolved from their original versions (some to baroque complexity, involving hundreds of free parameters), require catch-at-age data. Hence government laboratories, at least in developed countries, spend a large part of their budget on the routine acquisition and interpretation of catch and age-composition data.

Yet, single-species assessment models and the related policies have not served us particularly well, due to at least four broad problems. First, assessment results, although implying limitation on levels of fishing mortality which would have helped maintain stocks if implemented, have often been ignored, on the excuse that they were not 'precise enough' to use as evidence for economically painful restriction of fishing (the 'burden of proof' problem⁸⁶).

Second, the assessment methods have failed badly in a few important cases involving rapid stock declines, and in particular have led us to grossly underestimate the severity of the decline and the increasing ('depensatory') impacts of fishing during the decline⁸⁷.

Third, there has been insufficient attention in some cases to regulatory tactics: the assessments and models have provided reasonable overall targets for management (estimates of long-term sustainable harvest), but we have failed to implement and even develop effective short-term regulatory systems for achieving those targets⁸⁸.

Fourth, we have seen apparently severe violation of the assumptions usually made about 'compensatory responses' in recruitment to reduction in spawning population size. We have usually assumed that decreasing egg production will result in improving juvenile survival (compensation) so that recruitment (eggs $\times$ survival) will not fall off rapidly during a stock decline and will hence tend to stop the decline. Some stocks have shown recruitment failure after severe decline, possibly associated with changes in feeding interactions that are becoming known as 'cultivation/depensation' effects⁸⁹. According to this phenomenon, adult predatory fish (such as cod) can control the abundance of potential predators and competitors of their juvenile offspring, but this control lost when these predatory fish become scarce. This may well lead to alternate stable states of ecosystems, which has severe implications for fisheries management⁹⁰.

Jointly, these four broad problems imply a need to complement our single-species assessments by elements drawn from ecology, that is, to move towards ecosystem-based management. What this will consist of is not clearly established, although it is likely that, while retaining single-species models at its core, it will have to explicitly include trophic interaction between species⁹¹, habitat impacts of various gears⁵⁰, and a theory for dealing with the optimum placement and size of marine reserves (see main text). Ecosystem-based management will have to rely on the principles of, and lessons learnt from, single-species stock assessments, especially regarding the need to limit fishing mortality. It will certainly not be applicable in areas where effort or catch limits derived from single-species approaches cannot be implemented in the first place.

Throughout the 1950s and 1960s, this huge increase of global fishing effort led to an increase in catches (Fig. 1) so rapid that their trend exceeded human population growth, encouraging an entire generation of managers and politicians to believe that launching more boats would automatically lead to higher catches.

The first collapse with global repercussions was that of the Peruvian anchoveta in 1971–1972, which is often perceived as having been caused by an El Niño event. However, much of the available evidence, including actual catches (about 18 million tonnes²²) exceeding officially reported catches (12 million tonnes), suggest that overfishing was implicated as well. But attributing the collapse of the Peruvian anchoveta to 'environmental effects' allowed business as usual to continue and, in the mid-1970s, this led to the beginning of a decline in total catches from the North Atlantic. The declining trend accelerated in the late 1980s and early 1990s when most of the cod stocks off New England and eastern Canada collapsed, ending fishing traditions reaching back for centuries²³.

Despite these collapses, the global expansion of effort continued¹⁴ and trade in fish products intensified to the extent that they have now become some of the most globalized commodities, whose price increased much faster than the cost of living index²⁴. In 1996, FAO published a chronicle of global fisheries showing that a rapidly increasing fraction of world catches originate from stocks that are depleted or collapsed, that is, 'senescent' in FAO's parlance²⁵. Yet,

Box 2

Trophic levels as indicators of fisheries impacts

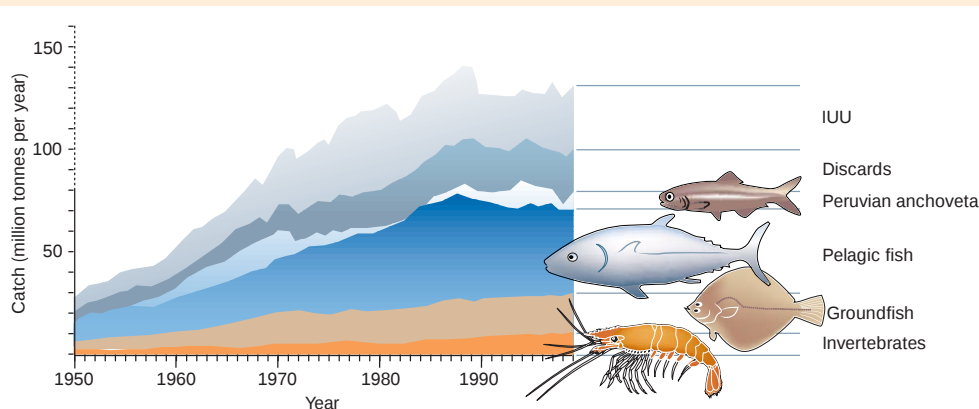
There are many ways ecosystems can be described, for example in terms of the information that is exchanged as their components interact, or in terms of size spectra. But perhaps the most straightforward way to describe ecosystems is in terms of the feeding interactions among their component species, which can be done by studying their stomach contents. A vast historical database of such published studies exists²⁷, which has enabled a number of useful generalizations to be made for ecosystem-based management of fisheries. One of these is that marine systems have herbivores (zooplankton) that are usually much smaller than the first-order carnivores (small fishes), which are themselves consumed by much larger piscivorous fishes, and so on. This is a significant difference from terrestrial systems, where, for example, wolves are smaller than the moose they prey on. Another generalization is that the organisms we have so far extracted from marine food webs have tended to play therein roles very different from those played by the terrestrial animals we consume. This can be shown in terms of their 'trophic level' (TL), defined as 1 + the mean TL of their prey.

Thus, in marine systems we have: algae at the bottom of the food web (TL = 1, by definition); herbivorous zooplankton feeding on the algae (TL = 2); large zooplankton or small fishes, feeding on the herbivorous zooplankton (TL = 3); large fishes (for example, cod, tuna and groupers) whose food tends to be a mixture of low- and high-TL organisms (TL = 3.5–4.5).

The mean TL of fisheries landings can be used as an index of sustainability in exploited marine ecosystems. Fisheries tend at first to remove large, slower-growing fishes, and thus reduce the mean TL of the fish remaining in an ecosystem. This eventually leads to declining trends of mean TL in the catches extracted from that ecosystem, a process now known as 'fishing down marine food webs'²⁹.

Declining TL is an effect that occurs within species as well as between species. Most fishes are hatched as tiny larvae that feed on herbivorous zooplankton. At this stage they have a TL of about 3, but this value increases with size, especially in piscivorous species. Because fisheries tend to reduce the size of the fish in an exploited stock, they also reduce their TL.

Figure 1 Estimated global fish landings 1950–1999. Figures for invertebrates, groundfish, pelagic fish and Peruvian anchoveta are from FAO catch statistics, with adjustment for over-reporting from China²⁶. Fish caught but then discarded were not included in the FAO landings; data relate to the early 1990s⁸³ were made proportional to the FAO landings for other periods. Other illegal, unreported or unregulated (IUU) catches⁶⁵ were estimated by identifying, for each 5-year block, the dominant jurisdiction and gear use (and hence incentive for IUU)⁸⁴; reported catches were then raised by the percentage of IUU in major fisheries for each 5-year block. The resulting estimates of IUU are very tentative (note dotted y-axis), and we consider that complementing landings statistics with more reliable estimates of discards and IUU is crucial for a transition to ecosystem-based management.



global catches seemed to continue, increasing through the 1990s according to official catch statistics. This surprising result was explained recently when massive over-reporting of marine fisheries catches by one single country, the People's Republic of China, was uncovered²⁶. Correcting for this showed that reported world fisheries landings have in fact been declining slowly since the late 1980s, by about 0.7 million tonnes per year.

Fisheries impact on ecosystem and biodiversity

The position within ecosystems of the fishes and invertebrates landed by fisheries can be expressed by their trophic levels, expressing the number of steps they are removed from the algae (occupying a trophic level of 1) that fuel marine food webs (Box 2). Most food fishes have trophic levels ranging from 3.0 to 4.5, that is, from sardines feeding on zooplankton to large cod or tuna feeding on miscellaneous fishes²⁷. Thus, the observed global decline of 0.05–0.10 trophic levels per decade in global fisheries landings (Fig. 2) is extremely worrisome, as it implies the gradual removal of large, long-lived fishes from the ecosystems of the world oceans. This is perhaps most clearly illustrated by a recent study in the North Atlantic showing that the biomass of predatory fishes (with a trophic level of 3.75 or more) declined by two-thirds through the second half to the twentieth century, even though this area was already severely depleted before the start of this time period²⁸.

It may be argued that so-called 'fishing down marine food webs' is both a good and an unavoidable thing, given a growing demand for fish²⁹. Indeed, the initial ecosystem reaction to the process may be a release from predation, where cascading effects may lead to increased catches³⁰. Such effects are, however, seldom observed in marine ecosystems^{31,32}, mainly because they do not function simply as a number of unconnected food chains. Rather, predators operate within finely meshed food webs, whose structure (which they help maintain) tends to support the production of their prey. Hence the concept of 'beneficial predation', where a predator may have a direct negative impact on its prey, but also an indirect positive effect, by consuming other predators and competitors of the prey³³ (and see Box 1). Thus, removing predators does not necessarily lead to more of their prey becoming available for humans. Instead, it leads to increases or outbursts of previously suppressed species, often invertebrates^{30,34,35}, some of which may be exploited (for example, squid or jellyfish, the latter a relatively new resource, exported to east Asia), and some outright noxious³⁶.

The principal, direct impact of fishing is that it reduces the abundance of target species. It has often been assumed that this does

not impose any direct threat of species extinction as marine fish generally are very fecund and the ocean expanse is wide³⁷. But the past few decades have witnessed a growing awareness that fishes can not only be severely depleted, but also be threatened with extinction through overexploitation³⁸. Among commercially important species, those particularly at risk are species that are highly valued, large and slow to mature, have limited geographical range, and/or have sporadic recruitment³⁹. There is actually little support, though, for the general assumption that the most highly fecund marine fish species are less susceptible to overexploitation; rather it seems that this perception is flawed⁴⁰. Fisheries may also change the evolutionary characteristics of populations by selectively removing the larger, fast-growing individuals, and one important research question is whether this induces irreversible changes in the gene pool⁴¹. Overall, this has implications for research, monitoring and management, and it points to the need for incorporating ecological consideration in fisheries management^{42,43}, as exemplified by the development of quantitative guidelines to avoid local extinctions⁴⁴.

Another worrisome aspect of fishing down marine food webs is that it involves a reduction of the number and length of pathways linking food fishes to the primary producers, and hence a simplification of the food webs. Diversified food webs allow predators to switch between prey as their abundance fluctuates⁴⁵, and hence to compensate for prey fluctuations induced by environmental fluctuations⁴⁶. Fisheries-induced food-web simplification, combined with the drastic fisheries-induced reduction in the number of year classes in predator populations^{47,48}, makes their reduced biomass strongly dependent of annual recruitment. This leads to increasing variability, and to lack of predictability in population sizes, and hence in predicted catches. The net effect is that it will increasingly look like environmental fluctuations impact strongly on fisheries resources, even where they originally did not. This resolves, if in a perverse way, the question of the relative importance of fisheries and environmental variability as the major driver for changes in the abundance of fisheries resources⁴⁹ (Fig. 3).

It seems unbelievable in retrospect, but there was a time when it was believed that bottom trawling had little detrimental impact, or even a beneficial impact, on the sea bottom that it 'ploughed'. Recent research shows that the ploughing analogy is inappropriate and that if an analogy is required, it should be that of clear cutting forests in the course of hunting deer. Indeed, the productivity of the benthic organisms at the base of food webs leading to food fishes is seriously impacted by bottom trawling⁵⁰, as is the survival of their juveniles when deprived of the biogenic bottom structure destroyed by that

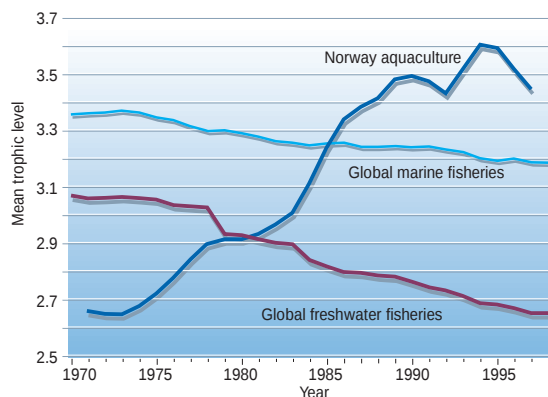


Figure 2 Fisheries, both marine and freshwater, are characterized by a decline of the mean trophic level in the landings, implying an increased reliance on organisms low in food webs (data from FishBase²⁷, with Peru/Chile excluded owing to the dominance of Peruvian anchoveta; see also Fig. 1). Freshwater fisheries have lower trophic level values overall, indicating an earlier onset of the 'fishing down' phenomenon²⁹. The trend is inverted in non-Asian aquaculture, whose production consists increasingly of piscivorous organisms, as illustrated here for Norway (a major producer, yet representative country)⁸⁵.

form of fishing⁵¹. Hence, given the extensive coverage of the world's shelf ecosystems by bottom trawling⁵², it is not surprising that generally longer-lived, demersal (bottom) fishes have tended to decline faster than shorter-lived, pelagic (open water) fishes, a trend also indicated by changes in the ratio of piscivorous (mainly demersal) to zooplanktivorous (mainly pelagic) fishes⁵³.

It is difficult to fully appreciate the extent of the changes to ecosystems that fishing has wrought, given shifting baselines as to what is considered a pristine ecosystem^{1,54} and continued reliance on single-species models (Box 1). These changes, often involving reductions of commercial fish biomasses to a few per cent of their pre-exploitation levels, prevent us taking much guidance from the concept of sustainability, understood as aiming to maintain what we have^{3,8}. Rather, the challenge is rebuilding the stocks in question.

Reducing fishing capacity

There is widespread awareness that increases in fishing-fleet capacity represent one of the main threats to the long-term survival of marine capture-fishery resources, and to the fisheries themselves^{55,56}. Reasons advanced for the overcapitalization of the world's fisheries include: the open-access nature of many fisheries⁵⁷; common-pool fisheries that are managed non-cooperatively^{58,59}; sole-ownership fisheries with high discount rates and/or high price-to-cost ratios⁶⁰; the increasing replacement of small-scale fishing vessels with larger ones⁵⁵; and the payment of subsidies by governments to fishers⁶¹, which generate 'profits' even when resources are overfished.

This literature shows that fishing overcapacity is likely to build up not only under open access⁶², but also under all forms of property regimes. Subsidies, which amount to US\$2.5 billion for the North Atlantic alone, exacerbate the problems arising from the open access and/or 'common pool' aspects of capture fisheries, including fisheries with full-fledged property rights^{61,63}.

Even subsidies used for vessel decommissioning schemes can have negative effects. In fact, decommissioning schemes can lead to the intended reduction in fleet size only if vessel owners are consistently caught by surprise by those offering this form of subsidy. As this is an unlikely proposition, decommissioning schemes often end up providing the collaterals that banks require to underwrite fleet modernizations. Additionally, in most cases, it is not the actual vessel that

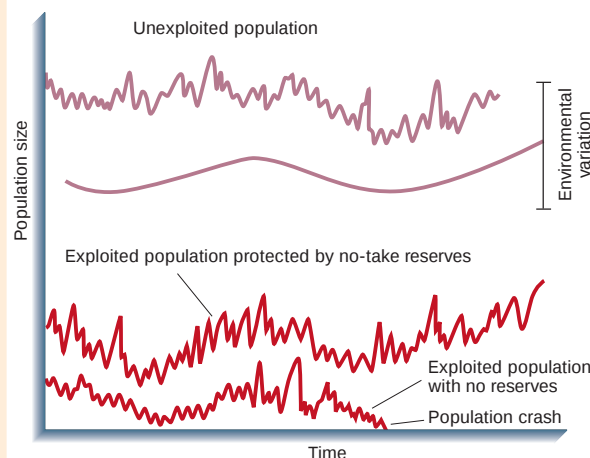


Figure 3 Schematic representation of the effects of some environmental variation on an unexploited, exploited but protected, and exploited but unprotected fish population. This illustrates how protection through a marine reserve (and/or stock rebuilding) can mitigate the effects of environmental fluctuations, including 'regime shifts'⁴⁹. (Graph from J. Jackson, personal communication.)

is retired, but its licence. This means that 'retired' vessels can still be used to catch species without quota (so-called 'under-utilized resources', which are often the prey of species for which there is a quota), or deployed along the coast of some developing country, the access to which may also be subsidized¹⁸. Clearly, the decommissioning schemes that will have to be implemented if we are ever to reduce overcapacity will have to address these deficiencies if they are not to end up, as most have so far, in fleet modernization and increased fishing mortality.

It is clear that a real, drastic reduction of overcapacity will have to occur if fisheries are to acquire some semblance of sustainability. The required reductions will have to be strong enough to reduce F by a factor of two or three in some areas, and even more in others. This must involve even greater decreases in f , because catches can be maintained in the face of dwindling biomasses by increasing q (and hence F ; see definitions above), even when nominal effort is constant. Indeed, this is the very reason behind the incessant technological innovation in fisheries, which now relies on global positioning systems and detailed maps of the sea bottom to seek out residual fish concentrations previously protected by rough terrain. This technological race, and the resulting increase in q , is also the reason why fishers often remain unaware of their own impacts on the resource they exploit and object so strongly to scientists' claims of reductions in biomass.

If fleet reduction is done properly, it should result in an increase in net benefits ('rent') from the resources, as predicted by the basic theory of bioeconomics⁶². This can be used, via taxation of the rent gained by the remaining fishers, to ease the transition of those who had to stop fishing. This would contrast with the present situation, where taxes from outside the fisheries sector are used, in form of subsidies, to maintain fishing at levels that are biologically unsustainable, and which ultimately lead to the depletion and collapse of the underlying resources.

Biological constraints to fisheries and aquaculture

Perhaps the strongest factor behind the politicians' use of tax money to subsidize non-sustainable, even destructive fisheries, and its tacit support by the public at large, is the notion that, somehow, the oceans will yield what we need — just because we need it. Indeed,

demand projections generated by national and international agencies largely reflect present consumption patterns, which by some means the oceans ought to help us maintain, even if the global human population were to double again. Although much of the deep ocean is indeed unexplored and 'mysterious', we know enough about ocean processes to realize that its productive capacity cannot keep up with an ever-increasing demand for fish.

Just as a tropical scientist might look at the impressive expanse of Canada and assume that this country has boundless potential for agricultural production, unaware that in reality only the thin sliver of land along its southern border (5%) is arable, we terrestrial aliens have assumed that the expanse and depths of the world's oceans will provide for us in the ways that its more familiar coastal fringes have. But this assumption is very wrong. Of the 363 million square kilometres of ocean on this planet, less than 7% — the continental shelves — are shallower than 200 m, and some of this shelf area is covered by ice. Shelves generate the biological production supporting over 90% of global fish catches, the rest consisting of tuna and other oceanic organisms that gather their food from the vast, desert-like expanse of the open oceans.

The overwhelming majority of shelves are now 'sheltered' within the exclusive economic zones (EEZ) of maritime countries, which also include all coral reefs and their fisheries (Box 3). According to the 1982 United Nations Convention on the Law of the Sea⁶⁴, any country that cannot fully utilize the fisheries resource of its EEZ must make this surplus available to the fleet of other countries. This, along with eagerness for foreign exchange, political pressure¹⁸ and illegal fishing⁶⁵, has led to all of the world's shelves being trawled for bottom fish, purse-seined for pelagic fishes and illuminated to attract and catch squid (to the extent that satellites can map the night time location of fishing fleets as well as that of cities). Overall, about 35% of the primary production on the world's shelves is required to sustain the fisheries⁶⁶, a figure similar to the human appropriation of terrestrial primary production⁶⁷.

The constraints to fisheries expansion that this implies, combined with the declining catches alluded to above, have led to suggestions that aquaculture should be able to bridge the gap between supply and demand. Indeed, the impressive recent growth of reported aquaculture is often cited as evidence of the potential of that sector to meet the growing demand for fish, or even to 'feed the world'.

Three lines of argument suggest that this is unlikely. The first is that the rapidly growing global production figures underlying this documented growth are driven to a large extent by the People's Republic of China, which reported 63% of world aquaculture production in 1998. But it is now known that China not only over-reports its marine fisheries catches, but also the production of many other sectors of its economy⁶⁸. Thus, there is no reason to believe that global aquaculture production in the past decades has risen as much as officially reported.

Second, modern aquaculture practices are largely unsustainable: they consume natural resources at a high rate and, because of their intensity, they are extremely vulnerable to the pollution and disease outbreaks they induce. Thus, shrimp aquaculture ventures are in many cases operated as slash-and-burn operations, leaving devastated coastal habitats and human communities in their wake^{69,70}.

Third, much of what is described as aquaculture, at least in Europe, North America and other parts of the developed world, consists of feedlot operations in which carnivorous fish (mainly salmon, but also various sea bass and other species) are fattened on a diet rich in fish meal and oil. The idea makes commercial sense, as the farmed fish fetch a much higher market price than the fish ground up for fish meal (even though they may consist of species that are consumed by people, such as herring, sardine or mackerels, forming the bulk of the pelagic fishes in Fig. 1). The point is that operations of this type, which are directed to wealthy consumers, use up much more fish flesh than they produce, and hence cannot replace capture fisheries, especially in developing countries, where very few can

Box 3

Sustainable coral reef fisheries: an oxymoron?

Globally, 75% of coral reefs occur in developing countries where human populations are still increasing rapidly. Although coral reefs account for only 0.1% of the world's ocean, their fisheries resources provide tens of millions of people with food and livelihood⁹². Yet, their food security, as well as other ecosystem functions they provide, is threatened by various human activities, many of which, including forest and land management, are unrelated to fishing⁹³.

It has often been assumed that the high levels of primary productivity reported for coral reefs imply high fisheries yields⁹⁴. However, the long-held notion that coral reef fishes are 'fast turnover' species, capable of high productivity, is being increasingly challenged⁹⁵. Yield estimates for coral reefs vary widely, ranging from 0.2 to over 40 tonnes km⁻² yr⁻¹ (ref. 96), depending on what is defined as coral reef area, and as coral reef fishes^{96,97}. Taking yields from the central part of this range (5–15 tonnes km⁻² yr⁻¹) and the most comprehensive reef-area estimate available⁹², we derive an estimate for total global annual yield of 1.4–4.2 million tonnes.

Although these estimates represent only 2–5% of global fisheries catches, they provide an important, almost irreplaceable, source of animal protein to the populations of many developing countries⁹⁶.

Clearly, maintaining the biodiversity that is a characteristic of healthy reefs is the key to maintaining sustainable reef fisheries. Yet coral reefs throughout the world are being degraded rapidly, especially in developing countries⁹³. Concerns regarding overexploitation of reef fisheries are widespread^{1,75,98}. The entry of new, non-traditional fishers into reef fisheries has led to intense competition and the use of destructive fishing implements, such as explosives and poisons, a process known as 'malthusian overfishing'²¹.

Another major problem is the growing international trade for live reef fish⁹⁹, often associated with mobile fleets using cyanide fishing, and targeting species that often have limited ranges of movements¹⁰⁰. This leads to serial depletion of large coral reef fishes, notably the humphead wrasse (*Cheilinus undulatus* Labridae), groupers (Serranidae) and snappers (Lutjanidae), and to reefs devastated by the cyanide applications.

These fisheries, which destroy the habitat of the species upon which they rely, are inherently unsustainable. It can be expected that they will have to cease operating within a few decades, that is, before warm surface waters and sea-level rise overcome what may be left of the world's coral reefs.

afford imported smoked salmon. Indeed, this form of aquaculture represents another source of pressure on wild fish populations⁷¹.

Perspectives

We believe the concept of sustainability upon which most quantitative fisheries management is based⁷² to be flawed, because there is little point in sustaining stocks whose biomass is but a small fraction of its value at the onset of industrial-scale fishing. Rebuilding of marine systems is needed, and we foresee a practical restoration ecology for the oceans that can take place alongside the extraction of marine resources for human food. Reconciling these apparently dissonant goals provides a major challenge for fisheries ecologists, for the public, for management agencies and for the fishing industry⁷³. It is important here to realize that there is no reason to expect marine resources to keep pace with the demand that will result from our growing population, and hopefully, growing incomes in now impoverished parts of the world, although we note that fisheries designed to be sustainable in a world of scarcity may be profitable.

We argued in the beginning of this review that whatever semblance of sustainability fisheries in the past might have had was due

to their inability to cover the entire range inhabited by the wildlife species that were exploited, which thus had natural reserves. We further argued that the models used traditionally to assess fisheries, and to set catch limits, tend to require explicit knowledge on stock status and total withdrawal from stocks, that is, knowledge that will inherently remain imprecise and error prone. We also showed that generally overcapitalized fisheries are leading, globally, to the gradual elimination of large, long-lived fishes from marine ecosystems, and their replacement by shorter-lived fishes and invertebrates, operating within food webs that are much simplified and lack their former 'buffering' capacity.

If these trends are to be reversed, a huge reduction of fishing effort involving effective decommissioning of a large fraction of the world's fishing fleet will have to be implemented, along with fisheries regulations incorporating a strong form of the precautionary principle. The conceptual elements required for this are in place, for example, in form of the FAO Code of Conduct for Responsible Fisheries⁷⁴, but the required political will has been lacking so far, an absence that is becoming more glaring as increasing numbers of fisheries collapse throughout the world, and catches continue to decline.

Given the high level of uncertainty facing the management of fisheries, which induced several collapses, it has been suggested by numerous authors that closing a part of the fishing grounds would prevent overexploitation by setting an upper limit on fishing mortality. Marine protected areas (MPAs), with no-take reserves at their core, combined with a strongly limited effort in the remaining fishable areas, have been shown to have positive effects in helping to rebuild depleted stocks^{75–77}. In most cases, the successful MPAs were used to protect rather sedentary species, rebuild their biomass, and eventually sustain the fishery outside the reserves by exporting juveniles or adults⁷⁵. Although migrating species would not benefit from the local reduction in fishing mortality caused by an MPA^{78,79}, the MPA would still help some of these species by rebuilding the complexity of their habitat destroyed by trawling, and thus decrease mortality of their juveniles⁸⁰. Enforcement of the no-take zones within MPAs would benefit from the application of high technology (for example, satellite monitoring of fishing vessels), presently used mainly to increase fishing pressure.

There is still much fear among fisheries scientists, especially in extra-tropical areas, that the export of fish from such reserves would not be sufficient to compensate for the loss of fishing ground⁸¹. Although we agree that marine reserves are no panacea, the present trends in fisheries, combined with the low degree of protection presently afforded (only 0.01% of the world's ocean is effectively protected), virtually guarantee that more fish stocks will collapse, and that these collapses will be attributed to environmental fluctuations or climate change (Fig. 3). Moreover, many exploited fish populations and eventually fish species will become extinct. MPAs that cover a representative set of marine habitats should help prevent this, just like forest and other natural terrestrial habitats have enabled the survival of wildlife species which agriculture would have otherwise rendered extinct.

Focused studies on the appropriate size and location of marine reserves and their combination into networks, given locale-specific oceanographic conditions, should therefore be supported. This will lead to the identification of reserve designs that would optimize export to adjacent fished areas, and which could thus be offered to the affected coastal and fisher communities, whose consent and support will be required to establish marine reserves and restructure the fisheries⁸. The general public could also be involved, through eco-labelling and other market-driven schemes, and through support for conservation-orientated non-government organizations, which can complement the activities of governmental regulatory agencies.

In conclusion, we think that the restoration of marine ecosystems to some state that existed in the past is a logical policy

goal⁸². There is still time to achieve this, and for our fisheries to be put on a path towards sustainability. □

doi:10.1038/nature01017

- Jackson, J. B. C. *et al.* Historical overfishing and the recent collapse of coastal ecosystems. *Science* **293**, 629–638 (2001).
- Orensanz, J. M. L., Armstrong, J., Armstrong, D. & Hilborn, R. Crustacean resources are vulnerable to serial depletion—the multifaceted decline of crab and shrimp fisheries in the Greater Gulf of Alaska. *Rev. Fish Biol. Fish.* **8**, 117–176 (1998).
- Ludwig, D., Hilborn, R. & Walters, C. Uncertainty, resource exploitation, and conservation: Lessons from history. *Science* **260**, 17–18 (1993).
- Rosenberg, A. A., Fogarty, M. J., Sissenwine, M. P., Beddington, J. R. & Shepherd, J. G. Achieving sustainable use of renewable resources. *Science* **262**, 828–829 (1993).
- Boyd, R. T. in *Handbook of American Indians: Northwest Coast* (ed. Suttles, W.) 135–148 (Smithsonian Institution, Washington DC, 1990).
- Bohnsack, J. A. (Subcommittee Chair) NOAA Tech. Memo. NMFS-SEFC-261 (National Oceanic and Atmospheric Agency, Miami, 1990).
- Yellen, J. E., Brooks, A. S., Cornelissen, E., Mehlman, M. J. & Stewart, K. A middle stone-age worked bone industry from Katanda, Upper Semliki Valley, Zaire. *Science* **268**, 553–556 (1995).
- Pitcher, T. J. Fisheries managed to rebuild ecosystems? Reconstructing the past to salvage the future. *Ecol. Appl.* **11**, 601–617 (2001).
- Wing, E. S. The sustainability of resources used by native Americans on four Caribbean islands. *Int. J. Osteoarchaeol.* **11**, 112–126 (2001).
- Cushing, D. H. *The Provident Sea* (Cambridge Univ. Press, Cambridge, 1987).
- Hardy, A. *The Open Sea* (Collins, London, 1956).
- Schaefer, M. B. Some aspects of the dynamics of populations important to the management of the commercial marine fisheries. *Bull. Inter-Am. Trop. Tuna Comm.* **1**, 27–56 (1954).
- Beverton, R. J. H. & Holt, S. J. *On the Dynamics of Exploited Fish Populations* (Chapman and Hall, London, 1957; Facsimile reprint 1993).
- Garcia, S. M. & Newton, C. in *Global Trends: Fisheries Management* (ed. Sissenwine, M. P.) Am. Fish. Soc. Symp. **20**, 3–27 (American Fisheries Society, Bethesda, MD, 1997).
- Mace, P. M. A new role for MSY in single-species and ecosystem approaches to fisheries stock assessment and management. *Fish. Fish.* **2**, 2–32 (2001).
- Clark, W. G., Hare, S. R., Parma, A. M., Sullivan, P. J. & Trumble, R. J. Decadal changes in growth and recruitment of Pacific halibut (*Hippoglossus stenolepis*). *Can. J. Fish. Aquat. Sci.* **56**, 242–252 (1999).
- Koslow, J. A. *et al.* Continental slope and deep-sea fisheries: implications for a fragile ecosystem. *ICES J. Mar. Sci.* **57**, 548–557 (1999).
- Kaczynski, V. M. & Fluharty, D. L. European policies in West Africa: who benefits from fisheries agreements. *Mar. Policy* **26**, 75–93 (2002).
- Silvestre, G. & Pauly, D. *Status and Management of Tropical Coastal Fisheries in Asia* (ICLARM, Makati City, Philippines, 1997).
- Thorpe, A. & Bennett, E. Globalisation and the sustainability of world fisheries: a view from Latin America. *Mar. Resource Econ.* **16**, 143–164 (2001).
- Pauly, D. *Small-scale Fisheries in the Tropics: Marginality, Marginalization, and Some Implications for Fisheries Management* (eds Pikitch, E. K., Huppert, D. D. & Sissenwine, M. P.) (American Fisheries Society, Bethesda, MD, 1997).
- Castillo, S. & Mendoza, J. in *The Peruvian Anchoveta and its Upwelling Ecosystem: Three Decades of Change* (eds Pauly, D. & Tsukayama, I.) ICLARM Stud. Rev. **15**, 109–116 (ICLARM, Manila, Philippines, 1987).
- Myers, R. A., Hutchings, J. A. & Barrowman, N. J. Why do fish stocks collapse? The example of cod in Atlantic Canada. *Ecol. Appl.* **7**, 91–106 (1997).
- Sumaila, U. R. in *Proc. Expo '98 Conf. Ocean Food Webs Econ. Product.*, Lisbon, 1–3 July 1998 (eds Pauly, D., Christensen, V. & Coelho, L.) ACP-EU Fish. Res. Rep. **5**, 87 (ACP-EU, Brussels, 1999).
- Grainger, R. J. R. & Garcia, S. M. *Chronicles of Marine Fishery Landings (1950–1994). Trend Analysis and Fisheries Potential* FAO Fish. Tech. Pap. **359** (Food and Agriculture Organization of the United Nations, Rome, 1996).
- Watson, R. & Pauly, D. Systematic distortions in world fisheries catch trends. *Nature* **424**, 534–536 (2001).
- Froese, R. & Pauly, D. (eds) *FishBase 2000: Concepts, Design and Data Sources* (distributed with four CD-ROMs; updates available at <http://www.fishbase.org>) (ICLARM, Los Baños, Philippines, 2000).
- Christensen, V. *et al.* in *Fisheries Impacts on North Atlantic Ecosystems: Models and Analyses* (eds Guénette, S., Christensen, V. & Pauly, D.) Fisheries Centre Res. Rep. **9**(4), 1–25 (also available at <http://www.fisheries.ubc.ca>) (Fisheries Centre, Univ. British Columbia, Vancouver, 2002).
- Pauly, D., Christensen, V., Dalsgaard, J., Froese, R. & Torres, F. Jr Fishing down marine food webs. *Science* **279**, 860–863 (1998).
- Daskalov, G. M. Overfishing drives a trophic cascade in the Black Sea. *Mar. Ecol. Prog. Ser.* **225**, 53–63 (2002).
- Pace, M. L., Cole, J. J., Carpenter, S. R. & Kitchell, J. F. Trophic cascades revealed in diverse ecosystems. *Trends Ecol. Evol.* **14**, 483–488 (1999).
- Pinnegar, J. K. *et al.* Trophic cascades in benthic marine ecosystems: lessons for fisheries and protected-area management. *Environ. Conserv.* **27**, 179–200 (2000).
- Ulanowicz, R. E. & Puccia, C. J. Mixed trophic impacts in ecosystems. *Coenos* **5**, 7–16. (1990).
- Parsons, T. R. in *Food Webs: Integration of Patterns and Dynamics* (eds Polis, G. A. & Winemiller, K. D.) 352–357 (Chapman and Hall, New York, 1996).
- Mills, C. E. Jellyfish blooms: are populations increasing globally in response to changing ocean conditions? *Hydrobiologia* **451**, 55–68 (2001).
- Van Dolah, F. M., Roelke, D. & Greene, R. M. Health and ecological impacts of harmful algal blooms: risk assessment needs. *Hum. Ecol. Risk Assess.* **7**, 1329–1345 (2001).
- Pitcher, T. J. A cover story: fisheries may drive stocks to extinction. *Rev. Fish Biol. Fish.* **8**, 367–370 (1998).
- Casey, J. M. & Myers, R. A. Near extinction of a large, widely distributed fish. *Science* **281**, 690–692 (1998).
- Sadovy, Y. The threat of fishing to highly fecund fishes. *J. Fish Biol.* **59**, 90–108 (2001).
- Hutchings, J. A. Collapse and recovery of marine fishes. *Nature* **406**, 882–885 (2000).
- Law, R. Fishing, selection, and phenotypic evolution. *ICES J. Mar. Sci.* **57**, 659–668 (2000).
- Gislason, H., Sinclair, M., Sainsbury, K. & O'Boyle, R. Symposium overview: incorporating ecosystem objectives within fisheries management. *ICES J. Mar. Sci.* **57**, 468–475 (2000).

43. Hutchings, J. A. Numerical assessment in the front seat, ecology and evolution in the back seat: Time to change drivers in fisheries and aquatic sciences? *Mar. Ecol. Prog. Ser.* **208**, 299–313 (2000).
44. Punt, A. E. Extinction of marine renewable resources: a demographic analysis. *Popul. Ecol.* **42**, 19–27 (2000).
45. Stephens, D. W. & Krebs, J. R. *Foraging Theory* (Princeton Univ. Press, Princeton, 1986).
46. Neutel, A.-M., Heesterbeek, J. A. P. & de Ruiter, P. C. Stability in real food webs: weak links in long loops. *Science* **296**, 1120–1123 (2002).
47. Longhurst, A. Cod: perhaps if we all stood back a bit? *Fish. Res.* **38**, 101–108 (1998).
48. Stergiou, K. I. Overfishing, tropicalization of fish stocks, uncertainty and ecosystem management: reshaping Okham's razor. *Fish. Res.* **55**, 1–9 (2002).
49. Steele, J. H. Regime shifts in marine ecosystems. *Ecol. Appl.* **8**(Suppl. 1), S33–S36 (1998).
50. Hall, S. J. *The Effects of Fisheries on Ecosystems and Communities* (Blackwell, Oxford, 1998).
51. Turner, S. J., Thrush, S. F., Hewitt, J. E., Cummings, V. J. & Funnell, G. Fishing impacts and the degradation or loss of habitat structure. *Fish. Mgmt Ecol.* **6**, 401–420 (1999).
52. Watling, L. & Norse, E. A. Disturbance of the seabed by mobile fishing gear: a comparison to forest clearcutting. *Conserv. Biol.* **12**, 1180–1197 (1998).
53. Caddy, J. F. & Garibaldi, L. Apparent changes in the trophic composition of world marine harvests: the perspective from the FAO capture database. *Ocean Coastal Mgmt* **43**, 615–655 (2000).
54. Pauly, D. Anecdotes and the shifting baseline syndrome of fisheries. *Trends Ecol. Evol.* **10**, 430 (1995).
55. Weber, P. Facing limits to oceanic fisheries: Part II: The social consequences. *Nat. Res. Forum* **19**, 39–46 (1995).
56. Mace, P. M. in *Developing and Sustaining World Fisheries Resources* (Proc. 2nd World Fish. Congr.) (eds Hancock, D. H., Smith, D. C. & Beumer, J.) 1–20 (CSIRO Publishing, Collingwood, VIC, 1997).
57. Gordon, H. S. The economic theory of a common property resource: the fishery. *J. Polit. Econ.* **62**, 124–142 (1954).
58. Munro, G. The optimal management of transboundary renewable resources. *Can. J. Econ.* **12**, 355–376 (1979).
59. Sumaila, U. R. Cooperative and non-cooperative exploitation of the Arcto-Norwegian cod stock in the Barents Sea. *Environ. Resource Econ.* **10**, 147–165 (1997).
60. Sumaila, U. R. & Bawumia, M. in *Fish Ethics: Justice in the Canadian Fisheries* (eds Coward, H., Ommer, R. & Pitcher, T. J.) 140–153. (Institute of Social and Economic Research, Memorial University, St John's, Newfoundland, 2000).
61. Munro, G. R. & Sumaila, U. R. in *Fisheries Centre Research Report 9(5)* (eds Pitcher, T. J., Sumaila, U. R. & Pauly, D.) 10–27 (also available at <http://www.fisheries.ubc.ca>) (Fisheries Centre, Univ. British Columbia, Vancouver, 2001).
62. Clark, C. W. *Mathematical Bioeconomics: The Optimal Management of Renewable Resources* (Wiley, New York, 1990).
63. Milazzo, M. World Bank Tech. Pap. No. 406 (World Bank, Washington DC, 1998).
64. United Nations. *The United Nations Convention on the Law of the Sea* (United Nations, New York, 1982).
65. Bray, K. FAO Rep. IUU/2000/6, 53 (Food and Agriculture Organization of the United Nations, Rome, 2000).
66. Pauly, D. & Christensen, V. Primary production required to sustain global fisheries. *Nature* **374**, 255–257 (1995).
67. Vitousek, P. M., Ehrlich, P. R. & Ehrlich, A. H. Human appropriation of the products of photosynthesis. *BioScience* **36**, 368–373 (1986).
68. Rawski, T. G. & Xiao, W. Roundtable on Chinese economic statistics: introduction. *China Econ. Rev.* **12**, 298–302 (2001).
69. Feigon, L. A harbinger of the problems confronting China's economy and environment: the great Chinese shrimp disaster of 1993. *J. Contemp. China* **9**, 323–332 (2000).
70. Pullin, R. S. V., Rosenthal, H. & Maclean, J. L. (eds) *Environment and Aquaculture in Developing Countries* (ICLARM, Manila, Philippines, 1993).
71. Naylor, R. L. *et al.* Effect of aquaculture on world fish supplies. *Nature* **405**, 1017–1024 (2000).
72. Sainsbury, K. J., Punt, A. E. & Smith, A. D. M. Design of operational management strategies for achieving fishery ecosystem objectives. *ICES J. Mar. Sci.* **57**, 731–741 (2000).
73. Cochrane, K. L. Reconciling sustainability, economic efficiency and equity in fisheries: the one that got away? *Fish Fish.* **1**, 3–21 (2000).
74. Edeson, W. R. Current legal development. The Code of Conduct for Responsible Fisheries: an introduction. *Int. J. Mar. Coast. Law* **11**, 233–238 (1996).
75. Roberts, C. M., Bohnsack, J. A., Gell, F., Hawkins, J. P. & Goodridge, R. Effects of marine reserves on adjacent fisheries. *Science* **294**, 1920–1923 (2001).
76. Mosquera, I., Côté, I. M., Jennings, S. & Reynolds, J. D. Conservation benefits of marine reserves for fish populations. *Anim. Conserv.* **3**, 321–332 (2000).
77. Murawski, S. A., Brown, R., Lai, H. L., Rago, P. R. & Hendrickson, L. Large-scale closed areas as a fishery management tool in temperate marine systems: the Georges Bank experience. *Bull. Mar. Sci.* **66**, 775–798 (2000).
78. Guénette, S., Pitcher, T. J. & Walters, C. J. The potential of marine reserves for the management of northern cod in Newfoundland. *Bull. Mar. Sci.* **66**, 831–852 (2000).
79. Lipcius, R. N., Seitz, R. D., Goldsborough, W. J., Montane, M. M. & Stockhausen, W. T. in *Spatial Processes and Management of Marine Populations* (ed. Witherell, D.) 643–666 (Alaska Sea Grant College Program, Anchorage, 2001).
80. Lindholm, J. B. I., Auster, P. J. & Kaufman, L. Habitat-mediated survivorship of 0-year Atlantic cod (*Gadus morhua*). *Mar. Ecol. Prog. Ser.* **180**, 247–255 (1999).
81. Tupper, M. K. *et al.* Marine reserves and fisheries management. *Science* **295**, 1233–1235 (2002).
82. Sumaila, U. R., Pitcher, T. J., Haggan, N. & Jones, R. in *Microbehavior and Macroresults: Proc. 10th Biannual Conf. Int. Inst. Fish. Econ. Trade, Corvallis, Oregon, 10–14 July 2000* (eds Johnston, R. S. & Shriver, A. L.) (International Institute of Fisheries Economics & Trade, Corvallis, OR, 2001).
83. Alverson, D. L., Freeberg, M. H., Murawski, S. A. & Pope, J. G. FAO Fish. Tech. Pap. 339, 233 (Food and Agriculture Organization of the United Nations, Rome, 1994).
84. Forrest, R., Pitcher, T., Watson, R., Valtýsson, H. & Guénette, S. in *Fisheries Impacts on North Atlantic Ecosystems: Evaluations and Policy Exploration* (ed. Pauly, D.) Fisheries Centre Rep. 5, 81–93 (also available at <http://www.fisheries.ubc.ca>) (Fisheries Centre, Univ. British Columbia, Vancouver, 2001).
85. Pauly, D., Tyedmers, P., Froese, R. & Liu, L. Y. Fishing down and farming up the food web. *Conserv. Biol.* **15**, 25 (2001).
86. Dayton, P. K. Ecology—reversal of the burden of proof in fisheries management. *Science* **279**, 821–822 (1998).
87. Walters, C. J. & Maguire, J. J. Lessons for stock assessment from the Northern Cod collapse. *Rev. Fish. Biol. Fish.* **6**, 125–137 (1996).
88. Perry, R. I., Walters, C. J. & Boutillier, J. A. A framework for providing scientific advice for the management of new and developing invertebrate fisheries. *Rev. Fish. Biol. Fish.* **9**, 125–150 (1999).
89. Walters, C. & Kitchell, J. F. Cultivation/densification effects on juvenile survival and recruitment: implications for the theory of fishing. *Can. J. Fish. Aquat. Sci.* **58**, 39–50 (2001).
90. Scheffer, M., Carpenter, S., Foley, J. A., Folke, C. & Walker, B. Catastrophic shifts in ecosystems. *Nature* **413**, 591–596 (2001).
91. Walters, C., Pauly, D., Christensen, V. & Kitchell, J. F. Representing density dependent consequences of life history strategies in aquatic ecosystems: EcoSim II. *Ecosystems* **3**, 70–83 (2000).
92. Spalding, M. D., Ravilious, C. & Green, E. P. *World Atlas of Coral Reefs* (Univ. California Press, Berkeley, 2001).
93. Roberts, C. M. *et al.* Marine biodiversity hotspots and conservation priorities for tropical reefs. *Science* **295**, 1280–1284 (2002).
94. Lewis, J. B. Processes of organic production on coral reefs. *Biol. Rev. Camb. Phil. Soc.* **52**, 305–347 (1977).
95. Choat, J. H. & Robertson, D. R. in *Coral Reef Fishes: Dynamics and Diversity in a Complex Ecosystem* (ed. Sale, P. F.) 57–80 (Academic, San Diego, 2002).
96. Russ, G. R. in *The Ecology of Fishes on Coral Reefs* (ed. Sale, P. F.) 601–635 (Academic, San Diego, 1991).
97. Spalding, M. D. & Grenfell, A. M. New estimates of global and regional coral reef areas. *Coral Reefs* **16**, 225–230 (1997).
98. Russ, G. R. in *Coral Reef Fishes: Dynamics and Diversity in a Complex Ecosystem* (ed. Sale, P. F.) 421–443 (Academic, San Diego, 2002).
99. Sadovy, Y. J. & Vincent, A. C. J. in *Coral Reef Fishes: Dynamics and Diversity in a Complex Ecosystem* (ed. Sale, P. F.) 391–420 (Academic, San Diego, 2002).
100. Zeller, D. C. Home range and activity patterns of the coral trout *Plectropomus leopardus* (Serranidae). *Mar. Ecol. Prog. Ser.* **154**, 65–77 (1997).

Acknowledgements

The work forms part of the Sea Around Us project at the Fisheries Centre, University of British Columbia, an activity initiated and funded by the Pew Charitable Trusts. D.P. and C.W. also acknowledge support from the Canadian National Scientific and Engineering Research Council. We thank G. Russ and M. L. 'Deng' Palomares for various suggestions that improved an earlier draft of our paper.

The present and future of the international wine industry

Linda F. Bisson*, Andrew L. Waterhouse*, Susan E. Ebeler*, M. Andrew Walker* & James T. Lapsley†

*Department of Viticulture and Enology and †UC Davis Extension, University of California, Davis, One Shields Avenue, Davis, California 95616-8749, USA

Wine production is both art and science, a blend of individual creativity and innovative technology. But wine production is also business, with economic factors driving manufacturing practices. To be successful in the modern marketplace, a winemaker must integrate the artistic and economic aspects of wine production, and possess a solid understanding of the intrinsic and extrinsic factors that underlie purchase motivation.

Wine is a unique commodity. Its production predates recorded history, as does the discovery of the healthful benefits of wine, now largely attributed to the antimicrobial activity of ethanol¹. Throughout antiquity the conversion of grapes into wine was considered a gift from the gods and the best wines were thus reserved for the elite of society. The image of wine as a beverage of the affluent persists even today. Wine was also one of the first commodities to be bartered by early civilizations engaged in international trade. Then, as now, the most successful wine producers were those who grasped market forces of supply and demand, and whose products met the prevailing definition of quality.

Today, wine is an integral component of the culture of many countries, a form of entertainment in others, and a libation of choice for advocates of its health benefits. Unlike many modern foods, wine's attractions rely not on bold consistent flavours, but upon a subtle array of shifting sensations that make its charm difficult to define. In essence, wine producers are selling a sensory experience to the consumer. Wine consumers in developed nations are typically prosperous, but wine is also consumed in impoverished areas where it is still safer to drink than the local water supply. Regardless of the region in which the wine is produced or the economic status of the consumer, all wines are expected to be pleasant experiences for the imbibor.

In past generations, the definition of quality was the preserve of the wine producer, and consumers who did not like a particular style of wine were often made to feel uncultured. But globalization and the accompanying rapid worldwide access to information has resulted in a more knowledgeable and empowered consumer with a more sophisticated understanding of product value and a discriminating demand for quality. The control of the definition of quality has thus shifted to the consumer. Success as a wine producer in the twenty-first century requires a thorough appreciation of human behaviour and product choice.

Wine is again a unique commodity in this respect. The intrinsic sensory aspect of wine taste and aroma are only one component in the modern consumer definition of quality. Extrinsic factors such as bottle and label design and the perceived artistic talents of the winemaker are equally important motivators of human preference in wine selection. In addition to a product that is enjoyable in all sensory aspects, consumers expect wines to be healthful and produced in an environmentally sustainable manner. In

the future, these last two factors will become increasingly important economic drivers of profitability.

These issues are complex, requiring producers to understand the latest developments in wide-ranging disciplines of science and technology. The present-day wine industry is focused on optimizing the attractiveness of the product within the bottle. In the future, the industry will need to go beyond this, paying more attention to the extrinsic factors motivating product choice, while ensuring that production remains cost-effective and economically sound.

The economics of wine production

In contrast to other types of crops, grapes can be grown in diverse climates and soils. Although scientifically still unproven, environmental stress is believed to improve the sensory characteristics of grapes and wine, resulting in a better product. The French concept of *terroir* states that the composition of grapes produced in a specific growing region will be influenced by the local environment, which will carry through to the wines of the area². This concept also includes as an element minimal intervention in modification of the growing environment so that the *terroir* may be evident. Thus, in contrast to other agricultural commodities, wine is marketed by the geographical location of production, and quality is associated with minimal vineyard inputs or manipulation.

Consumers expect wine from a particular region to possess unique qualities that differentiate it from other wines of the same varietal from other regions. This peculiarity of the industry is a great economic equalizer across the globe. It means that wines perceived to be of high quality can be produced anywhere. Indeed, quality wines are currently being produced on all six arable continents, and affluent as well as emerging nations are active in the international wine trade. The heightened tourism that accompanies the 'discovery' of a new wine-producing region is economically important to many countries. This 'value added' economic aspect of wine production is remarkable, and the main reason that many governments support strong research programmes in the development and improvement of their wine industries.

The wine industry is actually a composite of several individual economic or market segments. In the United States, roughly 70% of the market is comprised of the 'economy' wines, those that retail for less than US\$7 per 750-ml bottle. Wines range from economy to premium, ultra-premium and artisan, with wines in the latter two

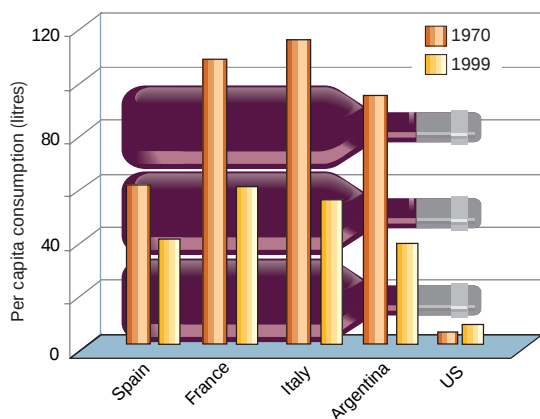


Figure 1 Changes in total consumption of wine by country over the past three decades. Data from refs 3, 4.

categories comprising only 2–3% of the market and commanding high prices. However, consumers have high expectations for product quality in all price categories. The actual sensory attributes desired may differ, but not the expectation of a pleasing sensory experience.

During the last third of the twentieth century the world wine market became significantly more competitive. Consumption declined in the traditional wine producing and consuming countries, while competition emerged from such 'New World' nations as the United States, Australia and Chile, and prosperous consumers chose quality rather than quantity in consumption. In 2001, France, Italy and Spain combined to produce slightly more than half of all the world's wine, but in the past 30 years their own per capita consumption has fallen 40–50% (Fig. 1), leading to an oversupply of 'Old World' wine^{3,4}. During this same period, US per capita wine consumption has almost doubled and, more important, US consumers have chosen to drink more expensive wine in a search for quality, a trend that seems to be true of European wine consumption as well. The New World producers have been quick to respond to global perceptions of quality, and have gained significant market share in the past 20 years, moving from 2 to 15% of the world export market (Fig. 2), largely at the expense of the European producers⁵.

Factors other than enhanced product quality have fuelled the increase in US consumption of wine. In 1991, a study by Serge Renaud coined the term 'French paradox' to describe the relationship between the high intake of fats in the French diet and the low incidence of coronary heart disease⁶. This well-turned phrase galvanized the attention of the media, the public and other scientists. Although there are many differences between the French and American diets, attention quickly focused on the disparity in wine consumption between the two countries. This was because of the long recognized benefit of moderate alcohol consumption^{1,7}. Renaud had no mechanistic explanation for the paradox, which led many to doubt its validity, but Kinsella and colleagues⁸ proposed that the natural antioxidant phenolic compounds of wine and fruits and vegetables of the Mediterranean diet might protect against heart disease. This conclusion was based partly upon a new theory advanced by Steinberg⁹ that linked oxidation in the blood to disease.

The power of the antioxidant hypothesis, coupled to the high visibility of Renaud's report and strong pressure from the public sector, promoted investigation into the chemical and biological activities of alcohol, dietary phenols and flavonoids. Today there is scientific evidence that moderate alcohol and/or wine consumption protects against the incidence of many diseases of modern society—cardiovascular disease, dietary cancers, ischaemic stroke, peripheral vascular disease, diabetes, hypertension, peptic ulcers, kidney stones and

macular degeneration—in addition to stimulating resistance to infection and retention of bone density^{10–13}. The benefits of antioxidants are more pronounced in red wines as these wines contain a higher phenolic content, but white wines also offer some benefit to the consumer.

The impact of the French paradox and the popularization of this study by the media had a pronounced impact on the international wine industry. Consumers were willing to pay more for a product with a perceived health benefit, while still expecting a satisfying sensory experience. This market remains highly competitive, as wines produced anywhere in the world may possess the same health-promoting effects. Some growing regions are even marketing their wine based on antioxidant content.

Oenology in the twenty-first century

Wine has long been considered an art form. In 1880, Robert Louis Stevenson coined the term 'bottled poetry' to describe the quest for perfection by wine producers¹⁴. As with all art forms, the term 'quality' is subjective and what one consumer considers to be attractive may be perceived as spoiled by another. The diversity of preferences has been both a blessing and a bane to the wine industry. Producers must develop a clear style by which to distinguish their product from competitors, but know that not all consumers or critics may find that style appealing. In contrast to other commodities, the region of production, the artistic reputation of the producer, and the conditions of production are important factors in the perceived value of wine¹⁵. For these reasons, it is important that the complex interplay of physiological, genetic and environmental factors that underpin human choice and preference be understood. This is the challenge to scientists and producers for the new century.

The completion of the human genome project together with advances made in the field of neurobiology of behaviour are providing crucial information on the basis of preference and the subjective definition of quality. It has been estimated that roughly 2% of the human genome is devoted to olfactory receptors¹⁶. Studies have suggested that many of these genes (60–70%) are non-functional or pseudogenes in most primates^{17,18}, but this still leaves hundreds of functional receptors. Although the neurobiology of olfaction and taste is less well studied than the non-chemical senses, interesting findings have emerged. For example, there is a strong connection between perception of an odour and emotion, be it pleasant or unpleasant¹⁹. Genetic factors such as variation in the aldehyde dehydrogenase 2 genotype in humans, which leads to adverse reactions to alcoholic beverages, have been shown to influence acquired preferences²⁰. In addition to genetic factors, perception is also clearly influenced by prior experiences and expectations.

Over the past few decades, the field of sensory science has made critical contributions to our understanding of the variables that influence and contribute to the sensory perception of foods and beverages. Initially, sensory analysis was used simply as a component of quality control—to make certain that a product did not contain any objectionable odours or flavours that would make it unpalatable to most, if not all, consumers. The field has developed into a sophisticated endeavour relying on the use of human tasters as analytical tools. This poses significant challenges given the diversity of human experiences and therefore of preferences for various aromas, and the complex aroma profiles comprising wide arrays of detectable scents and odours.

Like sensory analysis, chemical analysis was limited initially to detection of defect compounds present in high concentrations. However, sampling procedures and analytical tools that can detect trace compounds present at nanomolar concentrations, now make it possible to understand the subtle nuances associated with varietal wine flavour²¹. Understanding the chemical composition of wine is insufficient in the prediction of human preference, and recent efforts of flavour chemists have been focused on linking chemical and sensory measurements of flavour. In one such application, a trained human subject sniffs the effluent from a gas chromatogram. As the compounds elute from the column, the qualitative and quantitative

responses of the sniffer are recorded and related to the signal provided by a chemical detector. By incorporating the human sense of smell in the analytical process, gas chromatography–olfactometry can link the detection and quantification of odorants to their sensory impact in wine²². Complex chromatograms can be reduced to a small subset of compounds that have significant odour impact. When the subset of compounds is recombined, the aroma properties of the mixture will closely mimic the properties of the original wine^{23–25}.

There is also growing recognition that the volatility and release of flavour compounds can be altered by interaction with other matrix components (for example, the presence of sugar, ethanol, lipids and polyphenols)^{26–28}. These studies reinforce the idea that flavour perception is dynamic and the result of a complex pattern of chemical and physical interactions in the food and in the mouth, which trigger the brain's response to gustatory, trigeminal and olfactory stimuli. Statistical tools of multivariate analysis and artificial neural networks are being used to relate the chemical and sensory information to the subjective preference responses of consumers. Descriptive analysis is used to characterize wine flavour quantitatively²⁹. Using this technique, judges identify sensory attributes that differentiate among a group of wines and then evaluate the wines for the intensity of each individual attribute. Flavour profiles (Fig. 3) can then be drawn that visually compare the differences in wines¹⁵. Finally, the role that non-sensory characteristics such as pricing, winemaker reputation and label information play in influencing consumer preferences¹⁵ can be related to the chemical and sensory models developed. These analyses document not only the enormous complexity of individual perceptions and preferences^{15,21}, but also the complexity of the tools that will be needed in the future to understand the relationships among chemistry, perception, preference and behaviour.

Sustaining viticulture in the twenty-first century

As consumers become more aware of the vulnerability of our global environment, the demand for sound agricultural production practices is increasing. In the future, the perception of the producer as a conscientious environmental steward will be an important influence on the consumer's purchasing decision. This is due in part to the fact that the typical wine consumer is well educated and affluent. The wine industry has taken a leadership role through the formation of international associations of governments, scientists and producers to reach a consensus on the practices that should be allowed for products destined for the international marketplace. These associations are developing farming practice guidelines that limit the impact of site-development issues such as removal of native vegetation, erosion and water use. The farming protocols also address disease and pest control and encourage pesticide applications that have the lowest possible environmental impact while maintaining efficacy³⁰.

Although the global normalization of viticultural practices is some years away, the efforts underway are laying an important foundation for other agricultural industries. Controlling the introduction of new pests and diseases, the adaptation of existing pests to current control strategies, and managing pests and diseases with attention to increasing public concern over short- and long-term environmental impacts are the greatest challenges facing viticulture. The globalization of the wine industry has intensified the need for new solutions to these problems.

A devastating example of the consequences of inadvertent pest distribution was the introduction of grape phylloxera to Europe from North America in the mid-1850s. This root aphid was able to decimate the European *Vitis vinifera* wine and table grapes, which lacked the evolved resistance to its feeding found in the North American grape species³¹. The European economy of the late 1800s was based largely on the production and distribution of wine, and phylloxera had a crippling economic impact. The problem was solved by a massive rootstock breeding and evaluation effort using the North American grape species in hybrid combinations to address phylloxera resistance and combine tolerance to soils and ease of

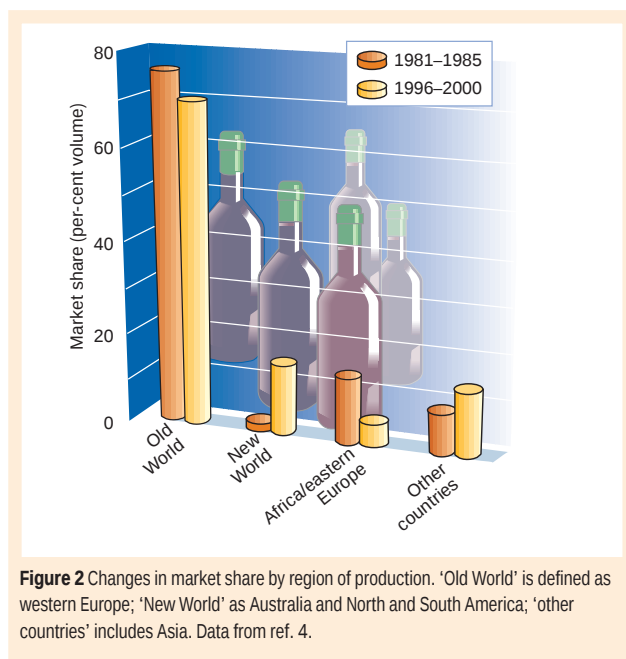


Figure 2 Changes in market share by region of production. 'Old World' is defined as western Europe; 'New World' as Australia and North and South America; 'other countries' includes Asia. Data from ref. 4.

propagation. This effort involved all wine producing countries and continues, on a smaller scale, to this day.

Although most grape-growing countries have quarantines on the movement of grape planting stock, the inadvertent movement of grape pests and diseases continues. The introduction of the glassy winged sharpshooter (*Homalodisca coagulata*) into California, most likely from Florida, has the potential to destroy large portions of California vineyards. This insect is an effective vector of Pierce's disease, a bacterial disease spread by xylem-feeding insects and native to the southern United States³². This introduction has led to mobilization of industry and government agencies and the allocation of millions of dollars for research and vector control (<http://www.cdffa.ca.gov/gwss/>). The primary focus of these efforts is control and, where possible, eradication of the vector. However, given that this insect is now established in southern California, its eventual movement throughout much of the state seems inevitable. In addition, the causal agent, the bacterium *Xylella fastidiosa*, is widely spread in a broad range of native and weedy plants, although limited in terms of its ability to cause disease by relatively inefficient native vectors³³. This combination of a new vector and common disease agent has heightened the need for new disease-resistant grape cultivars.

Disease-resistant grapes have been bred for the past 150 years. Because *V. vinifera* wine, table and raisin grapes do not possess any notable disease resistance, new cultivars must introgress or incorporate resistance genes that are most often found in the North American grape species. Resistance to a wide range of diseases is needed, but the production of wine is almost exclusively based on *V. vinifera* cultivars and even new hybrids within *V. vinifera* are poorly accepted owing to the industry's reliance on traditional and easily marketed classic wine grape cultivars. Classical breeding efforts to control Pierce's disease in table grapes are underway through introgression of resistance genes from southern US grape species into large-berried seedless *V. vinifera* table grapes. However, the production of wine grapes resistant to Pierce's disease will require the incorporation of resistance gene(s) via genetic engineering, owing to the reliance on classic *V. vinifera* wine grapes. There is an ongoing debate in the industry and by regulatory agencies as to the impact of a change in the genetic constitution of a grape cultivar and the ability to still refer to it as that cultivar. Varietal labelling is an important factor in wine marketing, and it is unclear how this will be impacted by the generation of genetically modified plants.

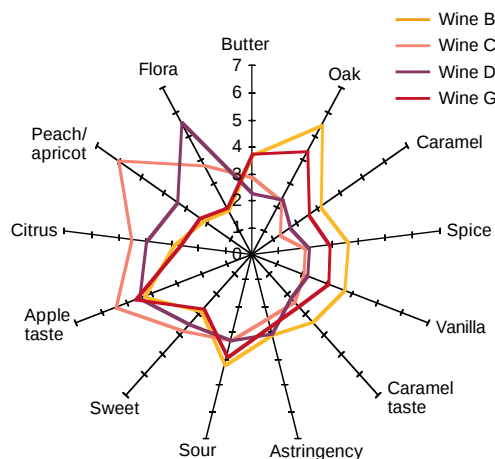


Figure 3 Flavour profiles of four inexpensive Chardonnay wines. Mean intensities for each term are plotted. The centre of the diagram corresponds to low intensity and the perimeter to high intensity. Data from ref. 15.

To meet the challenge of Pierce's disease and to develop greatly needed resistance to other pests and diseases, grape breeders and geneticists are beginning to direct much of their efforts towards mapping genes and genomics^{34–36}. These efforts are focused on developing a better understanding of the mechanisms, genetics and expression of pest and disease resistance and will lead to the development of *V. vinifera* grapes capable of being grown without pesticides or resistant rootstocks. The need for rapid progress is due not only to increased political and environmental pressure, but also to relatively unrestricted world travel and the internationalization and merger of the giants of the agricultural commodity industry, which virtually guarantee the spread of devastating pests and diseases to all corners of the world. Controlling these pests and diseases will require scientists to develop an exhaustive appreciation of native mechanisms of plant disease and pest resistance.

The future of the wine industry

The challenge to wine producers in this new century is daunting—to understand the fundamental motivation behind consumer choice and to produce wines of enhanced attractiveness while simultaneously developing and implementing sustainable production practices for both grape growing and wine making. Success in this endeavour necessitates determining the role of both intrinsic and extrinsic factors that underlie preference, perception and behaviour, and putting that information into practice.

As wine production has evolved from a cottage industry to global networks of consumer-aware producers, so has science. The challenges to scientists are the same as those to producers: to meet the demands of an inquiring public on issues that require an international and multidisciplinary approach. But beyond the wine sciences and the wine industry, the French paradox has had an equally dramatic impact on the fields of nutrition and preventive medicine and on food and beverage industries in general. The paradox has heightened consumer interest in the role of diet in health and health promotion, an interest society, as a whole would be wise to exploit. Sensory scientists are joining forces with neurobiologists to understand the allure of excess calorie consumption and to design a better yet equally attractive human diet. Wine will likely be a component of that diet—a lesson we could have learned from physicians throughout antiquity. This then is the future of food as gleaned from the experiences of the wine industry: a blending of nutritional benefits and sound environmental practices with the human perception of quality.

What then is the future of the industry? As we gain knowledge of the basic biology of human perception and flavour preferences, wines will become even more targeted to the genetic differences of the consumers. Consumer olfactory profiling will be common and used to guide production decisions as well as marketing of wines. Studies currently in progress will continue to document the healthful properties of wine. In addition, the industry will need to play a highly visible role in the promotion of sound and sustainable environmental stewardship, as this will be a strong motivating factor in the purchase of wines. The stakes of success in meeting consumer expectations are high, as the value-added aspects of enhanced tourism are undeniable and economically beneficial for the entire region. □

doi:10.1038/nature01018

- Lucia, S. P. A *History of Wine as Therapy* (Lippincott, Philadelphia, 1963).
- Laville, P. Le terroir, un concept indispensable à l'élaboration et à la protection des appellations d'origine comme la gestation des vignobles: le cas de la France. *Bull. OIV* **217**, 709–710 (1990).
- Protin, R. Situation de la viticulture dans le monde en 1970. *Bull. OIV* **44**, 110–1057 (1971).
- Anonymous. The state of viticulture in the world and the statistical information for 1999. *Bull. OIV* (Suppl.) 20–30 (2000).
- Aigrain, P. Conjuncture vitiviniculture mondiale. *Bull. OIV* **74**, 209–225 (2001).
- Renaud, S. & De Lorgeril, M. Wine, alcohol, platelets and the French paradox for coronary heart disease. *Lancet* **339**, 1523–1526 (1992).
- Marmot, M. G., Rose, G., Shipley, M. J. & Thomas, B. J. Alcohol and mortality: A U-shaped curve. *Lancet* **1**, 580–583 (1981).
- Kinsella, J. E., Frankel, E. N., German, J. B. & Kanner, J. Possible mechanisms for the protective role of antioxidants in wine and plant foods. *Food Tech.* **47**, 85–89 (1993).
- Steinberg, D. Modified forms of low-density lipoprotein and atherosclerosis. *J. Int. Med.* **233**, 227–232 (1993).
- German, B. G., Frankel, E. N., Waterhouse, A. L., Hansen, R. J. & Walzen, R. L. in *Wine: Nutritional and Therapeutic Benefits* (ed. Watkins, T. R.) 196–214 (Am. Chem. Soc., Washington DC, 1997).
- Gronbaek, M. et al. Type of alcohol consumed and mortality from all causes, coronary heart disease, and cancer. *Ann. Int. Med.* **133**, 411–419 (2000).
- Nijveldt, R. J. et al. Flavonoids: a review of probable mechanisms of action and potential applications. *Am. J. Clin. Nutr.* **74**, 418–425 (2001).
- De Lorimer, A. A. Alcohol, wine, and health. *Am. J. Surg.* **180**, 357–361 (2000). [Also available at <http://www.medicalfriendsofwinet.org/alcoholwine.htm> (2001).]
- Stevenson, R. L. *The Amateur Emigrant and The Silverado Squatters* (Scribner's Sons, New York, 1923).
- Yegge, J. & Noble, A. C. in *Proc. ASEV 50th Anniv. Annu. Meet.* (American Society for Enology and Viticulture, Davis, 2000).
- Firestein, S. How the olfactory system makes sense of scents. *Nature* **413**, 211–218 (2001).
- Rouquier, S., Blancher, A. & Giorgi, D. The olfactory receptor gene repertoire in primates and mouse: evidence for reduction of the functional fraction in primates. *Proc. Natl Acad. Sci. USA* **97**, 2870–2874 (2000).
- Sosinsky, A., Glusman, G. & Lancet, D. The genomic structure of human olfactory genes. *Genomics* **70**, 49–61 (2000).
- Brand, G., Millot, J.-L. & Henquell, D. Complexity of olfactory lateralization processes revealed by functional imaging: a review. *Neurosci. Biobehav. Rev.* **25**, 159–166 (2001).
- Ishibashi, T., Harada, S., Fugii, C., Taguchi, A. & Ishii, T. Relationship between ALDH2 genotypes and choice of alcoholic beverages. *Jpn. J. Alc. Stud. Drug Dep.* **34**, 117–129 (1999).
- Ebeler, S. E. in *Flavor Chemistry: 30 Years of Progress* (eds Teranishi, R., Wick, E. L. & Horstein, I.) 409–421 (Kluwer Academic/Plenum, New York, 1999).
- Acree, T. E. GC/Olfactometry: GC with a sense of smell. *Anal. Chem.* **69**, 170A–175A (1997).
- Guth, H. Identification of character impact odorants of different wine varieties. *J. Agric. Food Chem.* **45**, 3022–3026 (1997).
- Guth, H. Quantitation and sensory studies of character impact odorants of different white wine varieties. *J. Agric. Food Chem.* **45**, 3027–3032 (1997).
- Guth, H. in *Chemistry of Wine Flavor* (eds Waterhouse, A. L. & Ebeler, S. E.) 39–52 (Am. Chem. Soc., Washington DC, 1998).
- Voilley, A. & Lubbers, S. in *Chemistry of Wine Flavor* (eds Waterhouse, A. L. & Ebeler, S. E.) 217–229 (Am. Chem. Soc., Washington, DC, 1998).
- Dufour, C. & Bayonove, C. L. Interactions between wine polyphenols and aroma substances. An insight at the molecular level. *J. Agric. Food Chem.* **4**, 678–684 (1999).
- Jung, D.-M., De Ropp, J. S. & Ebeler, S. E. Study of interactions between food phenolics and aromatic flavors using one and two dimension 1H NMR spectroscopy. *J. Agric. Food Chem.* **48**, 407–412 (2000).
- Noble, A. C., Flath, R. A. & Forrey, R. Wine headspace analysis. Reproducibility and application to varietal classification. *J. Agric. Food Chem.* **28**, 346–353 (1980).
- Ohmart, C. P. & Matthiasson, S. K. *Lodi Winegrower's Workbook: A Self-Assessment of Integrated Farming Practice* <http://www.lodiwine.com/winegrowersworkbook1.shtml> (2000).
- Granett, J., Walker, M. A., Kocsis, L. & Omer, A. D. Biology and management of grape phylloxera. *Annu. Rev. Entomol.* **46**, 387–412 (2001).
- Hopkins, D. L. *Xylella fastidiosa*: a xylem-limited bacterial pathogen of plants. *Annu. Rev. Phytopathol.* **27**, 271–290 (1989).
- Purcell, A. H. & Saunders, S. R. Fate of Pierce's disease strains of *Xylella fastidiosa* in common riparian plants in California. *Plant Dis.* **83**, 825–830 (1999).
- Lodhi, M. A., Daly, M. J., Ye, G.-N., Weeden, N. F. & Reisch, B. I. A molecular marker based linkage map of Vitis. *Genome* **38**, 786–794 (1995).
- Paquet, J., Bouquet, A., This, P. & Adam-Blondin, A.-F. Establishment of a local map of AFLP markers around the powdery mildew resistance gene *Run1* in grapevine and assessment of their usefulness for marker assisted selection. *Theor. Appl. Genet.* **103**, 1201–1210 (2001).
- Donald, T. M. et al. Identification of resistance gene analogs linked to a powdery mildew resistance locus in grapevine. *Theor. Appl. Genet.* **104**, 610–618 (2002).

Evolution, consequences and future of plant and animal domestication

Jared Diamond

Department of Physiology, University of California Medical School, Los Angeles, California 90095-1751, USA
(e-mail: jdiamond@mednet.ucla.edu)

Domestication interests us as the most momentous change in Holocene human history. Why did it operate on so few wild species, in so few geographic areas? Why did people adopt it at all, why did they adopt it when they did, and how did it spread? The answers to these questions determined the remaking of the modern world, as farmers spread at the expense of hunter-gatherers and of other farmers.

Plant and animal domestication is the most important development in the past 13,000 years of human history. It interests all of us, scientists and non-scientists alike, because it provides most of our food today, it was prerequisite to the rise of civilization, and it transformed global demography. Because domestication ultimately yielded agents of conquest (for example, guns, germs and steel) but arose in only a few areas of the world, and in certain of those areas earlier than in others, the peoples who through biogeographic luck first acquired domesticates acquired enormous advantages over other peoples and expanded. As a result of those replacements, about 88% of all humans alive today speak some language belonging to one or another of a mere seven language families confined in the early Holocene to two small areas of Eurasia that happened to become the earliest centres of domestication — the Fertile Crescent and parts of China. Through that head start, the inhabitants of those two areas spread their languages and genes over much of the rest of the world. Those localized origins of domestication ultimately explain why this international journal of science is published in an Indo-European language rather than in Basque, Swahili, Quechua or Pitjantjatjara.

Much of this review is devoted to domestication itself: its origins, the biological changes involved, its surprising restriction to so few species, the restriction of its geographic origins to so few homelands, and its subsequent geographic expansion from those homelands. I then discuss the consequences of domestication for human societies, the origins of human infectious diseases, expansions of agricultural populations, and human evolution. After posing the unresolved questions that I would most like to see answered, I conclude by speculating about possible future domestications of plants and animals, and of ourselves. By a domesticate, I mean a species bred in captivity and thereby modified from its wild ancestors in ways making it more useful to humans who control its reproduction and (in the case of animals) its food supply. Domestication is thus distinct from mere taming of wild-born animals. Hannibal's African war elephants were, and modern Asian work elephants still are, just tamed wild individuals, not individuals of a genetically distinct population born and reared in captivity.

In 1997 I summarized available information about domestication and its consequences for human history in a book¹. Since then, new details have continued to

accumulate, and unanswered questions have come into sharper focus. Sources for statements not specifically referenced will generally be found in refs 1–9.

The past of domestication

Our 'decision' to domesticate

The question "why farm?" strikes most of us modern humans as silly. Of course it is better to grow wheat and cows than to forage for roots and snails. But in reality, that perspective is flawed by hindsight. Food production could not possibly have arisen through a conscious decision, because the world's first farmers had around them no model of farming to observe, hence they could not have known that there was a goal of domestication to strive for, and could not have guessed the consequences that domestication would bring for them. If they had actually foreseen the consequences, they would surely have outlawed the first steps towards domestication, because the archaeological and ethnographic record throughout the world shows that the transition from hunting and gathering to farming eventually resulted in more work, lower adult stature, worse nutritional condition and heavier disease burdens^{10,11}. The only peoples who could make a conscious choice about becoming farmers were hunter-gatherers living adjacent to the first farming communities, and they generally disliked what they saw and rejected farming, for the good reasons just mentioned and others.

Instead, the origins of domestication involved unforeseen consequences of two sets of changes — changes in plants and animals, and changes in human behaviour. As initially recognized by Darwin¹², and elaborated by Rindos¹³, many of the differences between domestic plants and their wild ancestors evolved as consequences of wild plants being selected, gathered and brought back to camp by hunter-gatherers, while the roots of animal domestication included the ubiquitous tendency of all peoples to try to tame or manage wild animals (including such unlikely candidates as ospreys, hyenas and grizzly bears). Although humans had been manipulating wild plants and animals for a long time, hunter-gatherer behaviour began to change at the end of the Pleistocene because of increasingly unpredictable climate, decreases in big-game species that were hunters' first-choice prey, and increasing human occupation of available habitats^{14,15}. To decrease the risk of unpredictable variation in food supply, people broadened their diets (the so-called broad-spectrum revolution) to second- and third-choice foods, which included more small

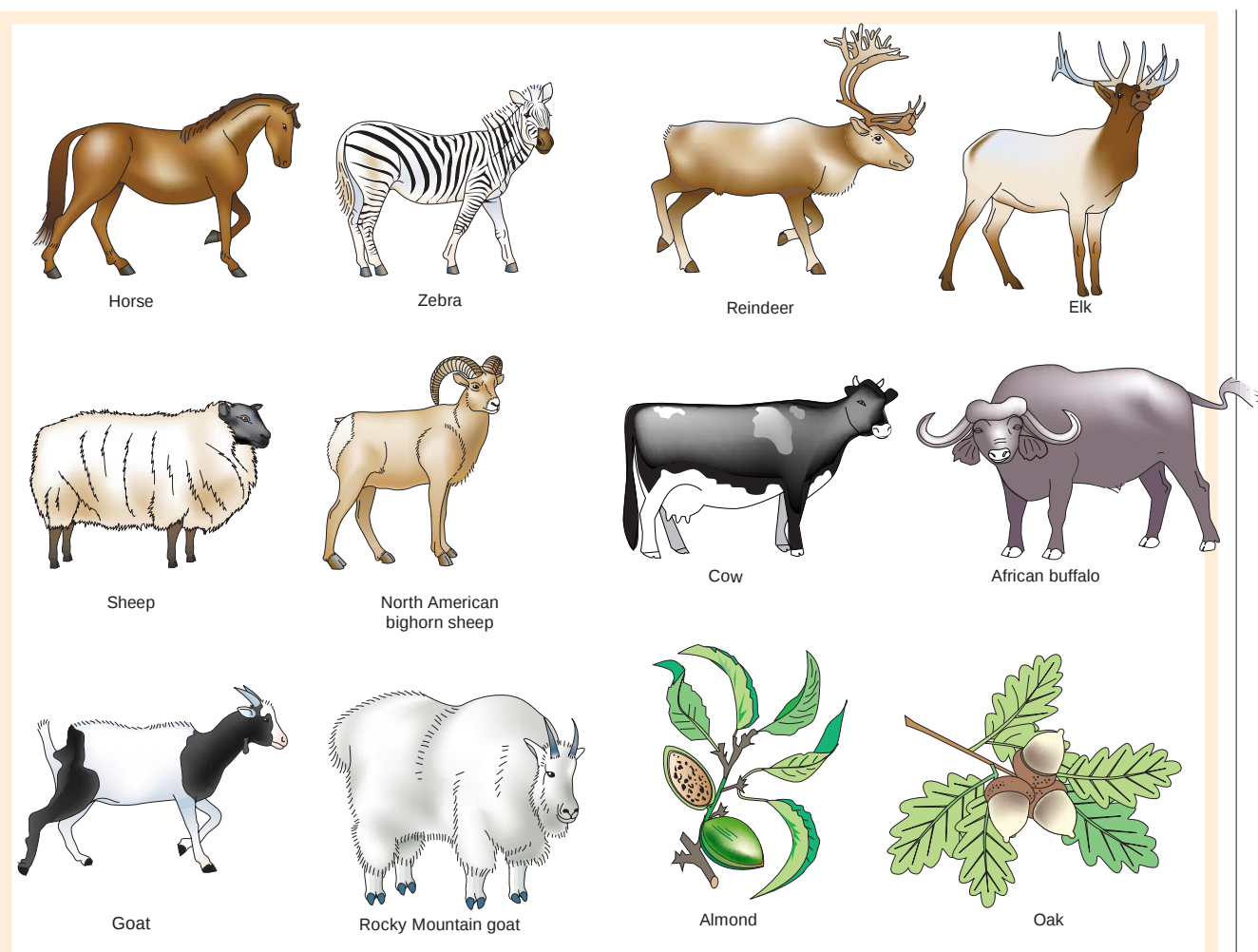


Figure 1 Comparisons of domesticated wild species (left of each pair) and their never-domesticated close relatives (right) reveal the subtle factors that can derail domestication.

game, plus plant foods requiring much preparation, such as grinding, leaching and soaking^{14,16}. Eventually, people transported some wild plants (such as wild cereals) from their natural habitats to more productive habitats and began intentional cultivation¹⁷.

The emerging agricultural lifestyle had to compete with the established hunter–gatherer lifestyle. Once domestication began to arise, the changes of plants and animals that followed automatically under domestication, and the competitive advantages that domestication conveyed upon the first farmers (despite their small stature and poor health), made the transition from the hunter–gatherer lifestyle to food production autocatalytic — but the speed of that transition varied considerably among regions^{18,19}. Thus, the real question about the origins of agriculture, which I consider below, is: why did food production eventually outcompete the hunter–gatherer lifestyle over almost the whole world, at the particular times and places that it did, but not at earlier times and other places?

Changes of wild species under domestication

These changes are particularly well understood for southwest Asia's Fertile Crescent, the site of domestication that was earliest in the world and that yielded what are still the world's most valuable domestic plant and animal species. For most species domesticated there, the wild ancestor and its wild geographic range have been identified, its relation to the domesticated proven by genetic and chromosomal studies, its changes under domestication delineated (often at the gene level), those changes traced in successive layers of

the archaeological record, and the approximate time and place of its domestication identified⁹.

For example, wild wheats and barley bear their seeds on top of a stalk that spontaneously shatters, dropping the seeds to the ground where they can germinate (but where they also become difficult for humans to gather). An occasional single-gene mutation that prevents shattering is lethal in the wild (because the seeds fail to drop), but conveniently concentrates the seeds for human gatherers. Once people started harvesting those wild cereal seeds, bringing them back to camp, accidentally spilling some, and eventually planting others, seeds with a non-shattering mutation became unconsciously selected for rather than against^{9,17}.

Individual wild animals also vary in traits affecting their desirability to humans. Chickens were selected to be larger, wild cattle (aurochs) to be smaller, and sheep to lose their bristly outer hairs (the kemp) and not to shed their soft inner hairs (the wool). Most domestic animals, including even recently domesticated trout²⁰, have smaller brains and less acute sense organs than do their wild ancestors. Good brains and keen eyes are essential to survival in the wild, but represent a quantitatively important waste of energy in the barnyard, as far as humans are concerned^{3,21}.

Especially instructive are cases in which the same ancestral species became selected under domestication for alternative purposes, resulting in very different-appearing breeds or crops. For instance, dogs were variously selected to kill wolves, dig out rats, race, be eaten, or be cuddled in our laps. What naive zoologist glancing at

wolfhounds, terriers, greyhounds, Mexican hairless dogs and chihuahuas would even guess them to belong to the same species? Similarly, cabbage (*Brassica oleracea*) was variously selected for its leaves (cabbage and kale), stems (kohlrabi), flower shoots (broccoli and cauliflower) and buds (brussels sprouts).

Why so few wild species were domesticated

The wild animal species that most plausibly could have yielded valuable domesticates were large terrestrial mammalian herbivores and omnivores, of which the world holds 148 species weighing 45 kg or more (Table 9.2 of ref. 1). Yet only 14 of those 148 species were actually domesticated (Table 9.1 of ref. 1), prompting us to ask what prevented domestication of the other 134 species? Similarly, world-wide there are about 200,000 wild species of higher plants, of which only about 100 yielded valuable domesticates. Especially surprising are the many cases in which only one of a closely related group of species became domesticated. For example, horses and donkeys were domesticated, but none of the four zebra species congeneric and able to interbreed with them^{3,22}.

The key question concerning this selectivity of domestication is as follows: in the cases of all those species never domesticated, did the difficulty lie with the species itself, or with the people indigenous to the area to which the species was native? For instance, is the abundance of large wild mammals the reason why no mammal species was ever domesticated in subequatorial Africa, making domestication superfluous for Africans? If that explanation were correct, then African people should also have ignored Eurasian domestic mammals when those were finally introduced to Africa, and European animal breeders on arriving in Africa should have succeeded in domesticating some African wild mammals, but both of those predictions are refuted by the actual course of history.

Six independent lines of evidence¹ converge to prove that, in most cases, the obstacle lay with the species itself, not with the local people: the rapid acceptance of introduced Eurasian domesticates by non-Eurasian peoples; the rapid ancient domestication of the most valuable wild species; the repeated independent domestications of many of them; the failure of even modern European plant and animal breeders to add significantly to our short list of valuable domesticates; ancient discoveries of the value of thousands of species that were regularly harvested in the wild but that never became domesticated; and the identification of the particular reasons preventing the domestication of many of those species.

Comparisons of domesticated wild species with never-domesticated close relatives illustrate the subtle factors that can derail domestication¹ (Fig. 1). For example, it is initially surprising that oak trees, the most important wild food plant in many parts of Eurasia and North America, were never domesticated. Like wild almonds, acorns of most individual wild oaks contain bitter poisons, with occasional non-poisonous mutant trees preferred by human foragers. However, the non-poisonous condition is controlled by a single dominant gene in almonds but polygenically in oaks, so that offspring of the occasional non-poisonous individuals are often non-poisonous in almonds but rarely so in oaks, preventing selection of edible oak varieties to this day. A second example is provided by the European horse breeders who settled in South Africa in the 1600s and — like African herders for previous millennia — tried to domesticate zebras. They gave up after several centuries for two reasons. First, zebras are incurably vicious, have the bad habit of biting a handler and not letting go until the handler is dead, and thereby injure more zoo-keepers each year than do tigers. Second, zebras have better peripheral vision than horses, making them impossible even for professional rodeo cowboys to lasso (they see the rope coming and flick away their head).

Among wild mammal species that were never domesticated, the six main obstacles proved to be a diet not easily supplied by humans (hence no domestic anteaters), slow growth rate and long birth spacing (for example, elephants and gorillas), nasty disposition (grizzly bears and rhinoceroses), reluctance to breed in captivity (pandas and

cheetahs), lack of follow-the-leader dominance hierarchies (bighorn sheep and antelope), and tendency to panic in enclosures or when faced with predators (gazelles and deer, except reindeer). Many species passed five of these six tests but were still not domesticated, because they failed a sixth test. Conclusions about non-domesticability from the fact of non-domestication are not circular, because these six obstacles can be assessed independently.

Why there were so few homelands of agriculture

Food production bestowed on farmers enormous demographic, technological, political and military advantages over neighbouring hunter-gatherers. The history of the past 13,000 years consists of tales of hunter-gatherer societies becoming driven out, infected, conquered or exterminated by farming societies in every area of the world suitable for farming. One might therefore have naively anticipated that, in any part of the world, one or more of the local hunter-gatherer societies would have stumbled upon domestication, become farmers, and thereby outcompeted the other local hunter-gatherer societies. In fact, food production arose independently in at most nine areas of the world (Fertile Crescent, China, Mesoamerica, Andes/Amazonia, eastern United States, Sahel, tropical West Africa, Ethiopia and New Guinea).

The puzzle increases when one scrutinizes that list of homelands. One might again naively have expected the areas most productive for farming today to correspond, at least roughly, to the areas most productive in the past. In reality, the list of homelands and the list of breadbaskets of the modern world are almost mutually exclusive (Fig. 2). The latter list includes California, North America's Great Plains, Europe, the pampas of Argentina, the cape of southern Africa, the Indian subcontinent, Java and Australia's wheat belt. Because these areas are evidently so well suited to farming or herding today, why were they not so in the past?

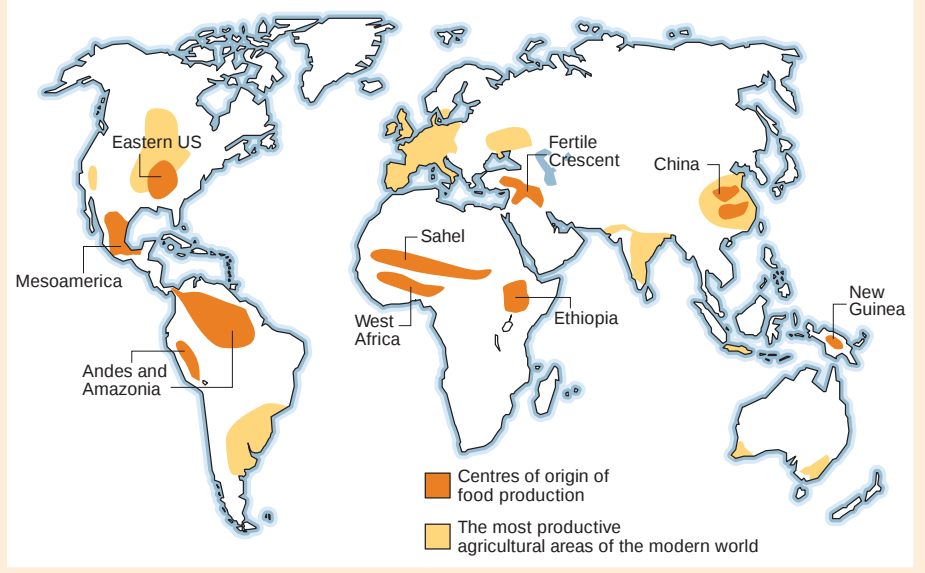
The explanation is that the homelands of agriculture were instead merely those regions to which the most numerous and most valuable domesticable wild plant and animal species were native. Only in those areas were incipient early farmers able to outcompete local hunter-gatherers. Once those locally available wild species had been domesticated and had spread outside the homelands, societies of homelands had no further advantage other than that of a head start, and they were eventually overtaken by societies of more fertile or climatically more favoured areas outside the homelands.

For instance, the Fertile Crescent of southwest Asia was home to wild wheats, barley, peas, sheep, goats, cows and pigs — a list that includes what are still the most valuable crops and livestock of the modern world. Hence hunter-gatherers of the Fertile Crescent domesticated those species and became the world's first farmers and herders, beginning around 8500 BC^{1,9,23}. That head start in food production led to them and their close neighbours also developing the world's first metal tools, writing, empires and professional armies. Those tools of conquest, and Fertile Crescent human genes, gradually spread west into Europe and North Africa and east into the western Indian subcontinent and central Asia. However, once those crops, livestock and human inventions had spread, Fertile Crescent societies possessed no other advantages. As all of those elements slowly spread northwest across Europe, farming and power also shifted northwest from the Fertile Crescent to areas where farming had never arisen independently — first to Greece, then to Italy, and finally to northwest Europe. Human societies of the Fertile Crescent inadvertently committed slow ecological suicide in a zone of low rainfall prone to deforestation, soil erosion and salinization.

The spread of food production

From the homelands of domestication, food production spread around the world in either of two ways. The much less common way was for hunter-gatherers outside the homelands to acquire crops or livestock from the homelands, enabling them to settle down as farmers or herders, as attested by archaeological evidence for substantial

Figure 2 Ancient and modern centres of agriculture. Ancient centres of origin of plant and animal domestication — the nine homelands of food production — are indicated by the orange-shaded areas on the map (based on Fig. 5.1 of ref. 1). The most agriculturally productive areas of the modern world, as judged by cereals and major staples, are indicated by the yellow-shaded areas. Note that there is almost no overlap between the areas highlighted, except that China appears on both distributions, and that the most productive areas of the central United States today approach areas of the eastern United States where domestication originated. The reason why the two distributions are so different is that agriculture arose in areas to which the wild ancestors of the most valuable domesticable crops and animals were native, but other areas proved much more productive when those valuable domesticates reached them.



continuity of material culture, and by genetic, linguistic and skeletal evidence of continuity of human populations. The clearest such example of local adoption of food production is in southern Africa, where around 2,000 years ago some Khoisan hunter-gatherers acquired Eurasian livestock (cattle, sheep and goats) arriving from the north and became herders (so-called Hottentots). Much more often, however, local hunter-gatherers had no opportunity to acquire crops and livestock before they were overrun or replaced by farmers expanding out of the homelands, exploiting their demographic, technological, political and military advantages over the hunter-gatherers.

Expansions of crops, livestock, and even people and technologies tended to occur more rapidly along east–west axes than along north–south axes¹ (Fig. 3). The reason is obvious: locations at the same latitude share identical day-lengths and seasonalities, often share similar climates, habitats and diseases, and hence require less evolutionary change or adaptation of domesticates, technologies and cultures than do locations at different latitudes. Examples include the rapid westwards and eastwards dispersal of wheat, horses, wheels and writing of western Asian origin, and the westwards dispersal of chickens, citrus and peaches of Chinese origin, along the east–west axis of Eurasia. This can be contrasted with the slow spread of Eurasian livestock and non-spread of Eurasian crops southwards along Africa's north–south axis²⁴, the slow spread of Mexican corn and the non-spread of Mexican writing and wheels and Andean llamas and potatoes along the Americas' north–south axis, and the slow spread of food production southwards along the north–south axis of the Indian subcontinent.

This is not to deny the existence of ecological barriers at the same latitude within Asia and North America, but the general pattern remains. Eurasia's east–west axis, and the resulting rapid enrichment of societies in each part of Eurasia by crops and technologies from other parts of Eurasia, became one of the main ultimate reasons why Eurasian peoples conquered Native American peoples, rather than visa versa. Eurasia's east–west axis also explains why there is much less evidence for multiple independent domestications of the same plant species (see below), and much more evidence for agriculturally driven language expansions, in Eurasia than in the Americas.

Consequences of domestication

Consequences for human societies

Beginning around 8500 BC, the transition from the hunter-gatherer lifestyle to food production enabled people to settle down next to

their permanent gardens, orchards and pastures, instead of migrating to follow seasonal shifts in wild food supplies. (Some hunter-gatherer societies in especially productive environments were also sedentary, but most were not). Food production was accompanied by a human population explosion that has continued unabated to this day, resulting from two separate factors. First, the sedentary lifestyle permitted shorter birth intervals. Nomadic hunter-gatherers had previously spaced out birth intervals at four years or more, because a mother shifting camp can carry only one infant or slow toddler. Second, plant and animal species that are edible to humans can be cultivated in much higher density in our gardens, orchards and pastures than in wild habitats.

Food production also led to an explosion of technology, because sedentary living permitted the accumulation of heavy technology (such as forges and printing presses) that nomadic hunter-gatherers could not carry, and because the storable food surpluses resulting from agriculture could be used to feed full-time craftspeople and inventors. By also feeding full-time kings, bureaucrats, nobles and soldiers, those food surpluses led to social stratification, political centralization and standing armies. All of these overwhelming advantages are what enabled farmers eventually to displace hunter-gatherers¹.

Evolution of epidemic infectious diseases

The main killers of humans since the advent of agriculture have been acute, highly infectious, epidemic diseases that are confined to humans and that either kill the victim quickly or, if the victim recovers, immunize him/her for life^{1,25–28}. Such diseases could not have existed before the origins of agriculture, because they can sustain themselves only in large dense populations that did not exist before agriculture, hence they are often termed 'crowd diseases'. The mystery of the origins of many of these diseases has been solved by molecular biological studies of recent decades, demonstrating that they evolved from similar epidemic diseases of our herd domestic animals with which we began to come into close contact 10,000 years ago. Thus, the evolution of these diseases depended on two separate roles of domestication: in creating much denser human populations, and in permitting much more frequent transmission of animal diseases from our domesticates than from hunted wild animals. For instance, measles and tuberculosis arose from diseases of cattle, influenza from a disease of pigs and ducks¹. An outstanding mystery remains the origins of smallpox: did it reach us from camels or from cattle?

Crowd diseases paradoxically became agents of conquest, because exposed individuals acquired immune resistance from childhood

exposure, and exposed populations gradually evolved genetic resistance, but unexposed populations had neither type of resistance. In practice, because 13 of our 14 large domestic mammals were Eurasian species, evolution of crowd diseases was concentrated in Eurasia, and the diseases became the most important agents by which Eurasian colonists expanding overseas killed indigenous peoples of the Americas, Australia, Pacific islands and southern Africa.

The agricultural expansions

Because some peoples acquired domesticates before other peoples could, and because domesticates conferred eventual advantages such as guns, germs and steel on the possessors, the history of the past 10,000 years has consisted of farmers replacing hunter–gatherers or less advanced farmers. These agricultural expansions, originating mainly from the nine homelands of agriculture, remade genetic and linguistic maps of the world (Table 18.2 of ref. 1). Among the most discussed (and often highly controversial) possible examples are the expansions of Bantu-speaking farmers out of tropical West Africa over subequatorial Africa²⁹, Austronesian-speaking farmers out of Taiwan over Island Southeast Asia³⁰, Fertile Crescent farmers over Europe^{31,32}, and Korean farmers over Japan³³.

Human genetic evolution

Domestication has been by far the most important cause of changes in human gene frequencies in the past 10,000 years. Among the mechanisms responsible are: the spread of human genes from the agricultural homelands; the evolution of genetic resistance factors (including the ABO blood groups) to our new crowd infectious diseases^{34,35}; the evolution of adult-persistent lactase in milk-consuming populations of northern Europe and several parts of Africa; the evolution of allozymes of alcohol metabolism permitting consumption of large quantities of nutritionally important beer in western Eurasia; and the evolution of adaptations to a diet higher in simple carbohydrates, saturated fats and (in modern times) calories and salt, and lower in fibre, complex carbohydrates, calcium and unsaturated fats, than the hunter–gatherer diet³⁶.

Unsolved questions

Among the host of unsolved questions, I focus here on six: what triggered the emergence of agriculture around 8500 BC and why did it not evolve earlier? Do crop and livestock species stem from a single domestication event or from multiple independent domestications? Can areas of food production be segregated into primary and secondary homelands, the latter describing areas where the arrival of primary homeland crops triggered local domestication? How did food production spread? Why were large domestic mammals predominantly Eurasian? And how can we gain a better understanding of the history of domestication of particular species?

Why then but not earlier?

The human lineage diverged from that of chimpanzees around 6,000,000 years ago. For the next 99.8% of our separate history, there was no agriculture, until it emerged independently in up to nine areas on four continents in the short span of 6,000 years between 8500 and 2500 BC. All of those nearly-simultaneous independent origins seem to be too much of a coincidence. What triggered agriculture repeatedly then, and why had it never arisen during the previous 6,000,000 years?

Posing the question in this way both understates and overstates the puzzle. It understates the puzzle, because there are not only up to nine independent trajectories of intensification that did culminate in agriculture, but also many other ones that didn't quite (or that hadn't yet at the time that European conquest aborted them). Areas of the world where hunter–gatherers in the Holocene developed increased population densities, complex material culture, in some cases pottery, and (some anthropologists argue) sedentary living and ranked societies with chiefs included Mesolithic Europe, Japan and

maritime Far East Asia, the North American high Arctic, the Pacific coast of northwest North America, interior California's oak woodlands, the California Channel Islands, the Calusa of Florida, the coast of Ecuador, and the Murray–Darling Basin of southeast Australia (for examples, see refs 37–39). But a similar intensification of hunter–gatherer societies also preceded the emergence of food production in its nine homelands; I suspect that the sole difference between the areas where people remained hunter–gatherers and the areas where food production evolved was that plant and animal species harvested in the latter but not the former areas included ones that automatically evolved domesticates, as already discussed. Thus, there were not just 5–9, but several dozen, independent trajectories of intensification in the Holocene.

On the other hand, my formulation of the question also overstates the puzzle. Only behaviourally modern *Homo sapiens* was biologically and mentally capable of the technological advances and foraging efficiency that resulted in intensified hunting and gathering, and (sometimes) in food production⁴⁰. But behaviourally modern *Homo sapiens* did not emerge until around 55,000–80,000 years ago (the exact date is debated), so we should say that the independent simultaneous emergences were not concentrated in the last 0.2% of hominid history, but 'only' in the last 15% of modern human history. Still, even that seems too concentrated a bout of simultaneous emergences to be coincidental. Was it just that the origins of behaviourally modern *Homo sapiens* set clocks ticking by chance at the same rate all over the globe? That strains credulity, especially as intensified hunter–gatherer economies failed to arise in more areas than the areas in which they did arise.

A possible explanation seems to me to derive from four developments in the Late Pleistocene that may indeed have driven the clock's ticking. First, improvements in human hunting skills and consequent depletion or extermination of large mammalian prey would have made the hunter–gatherer lifestyle less rewarding and less able to compete with food production. Second was the development of human technology to collect, process and store wild foods (such as wild cereals), without which subsequently exploiting the same food species as domesticates would have been impossible (that is, what is the point of sowing wheat if you have not yet determined how to reap, roast and store it?). The third development was the on-going competition between human societies, such that those societies with more effective technology at any moment prevailed over other societies. Fourth, the gradual rise in human population numbers through the Pleistocene required intensified food procurement to feed those larger populations.

Against that background of gradual change, a trigger that may have caused intensification and food production to emerge only after the end of the Pleistocene would have been the end-of-Pleistocene climate changes in temperature, rainfall and unpredictability. These changes could have triggered the broad-spectrum reduction in diet^{14–17}, and made agriculture possible in areas where it would have been impossible during the Ice Ages (for example, expanding Fertile Crescent woodland habitats with understories of wild cereals⁴¹). Once food production had thus begun, the autocatalytic nature of the many changes accompanying domestication (for example, more food stimulating population growth that required still more food) made the transition rapid. By this interpretation, the independent emergences of food production are no longer remarkably simultaneous — they could not have happened before the end of the Pleistocene (11000 BC), and after the end of the Pleistocene they occurred at very different times, ranging from about 8500 BC (in the Fertile Crescent) to about 2500 BC (eastern North America). Most of the links in this speculative hypothesis are in obvious need of testing.

Multiple versus single domestications

A long-standing question concerns whether each crop and livestock species stems from a single domestication event within a restricted

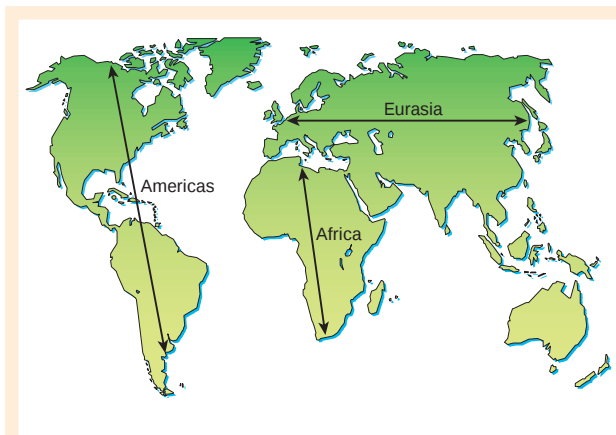


Figure 3 The continental major axis is oriented east–west for Eurasia but north–south for the Americas and Africa. The spread of food production tended to occur more rapidly along east–west axes than along north–south axes, mainly because locations at the same latitudes required less evolutionary change or adaptation of domesticates than did locations at different latitudes. Modified from Fig. 10.1 of ref. 1.

geographic area, or from multiple independent domestications at different sites. An accumulation of recent evidence suggests to me the following generalization: that the former interpretation applies to most major Eurasian crops, the latter interpretation to many New World crops and the major Eurasian livestock species.

Among New World crops, many are represented by distinct but related species in South America, Mesoamerica and the eastern United States, leaving no doubt that related species were domesticated independently in these areas (for example, beans, chenopods, chilli peppers, cotton, squashes, tobaccos and possibly amaranths). Multiple independent domestications are attested within the same species for the chilli pepper species *Capsicum annuum*, common bean *Phaseolus vulgaris*, lima bean *Phaseolus lunatus* and squash species *Cucurbita pepo*^{4,7,8,42}. Conversely, the eight crops that founded Fertile Crescent agriculture, with the possible exception of barley, each seem to derive from only a single domestication event^{5,9,43–45}.

Evidence for separate independent domestications in western and more eastern parts of Eurasia are now available for all of the ‘big five’ domesticated mammals (cow, sheep, goat, pig and horse), plus one of the ‘minor nine’ (water buffalo)^{46–54}. For example, cows were domesticated independently in the Fertile Crescent (yielding modern humpless cows), in the Indian subcontinent (yielding modern humped Zebu cows) and in North Africa^{46,50,53}.

I suggest the following hypothesis to explain predominantly single domestications of Fertile Crescent founder crops, but multiple domestications of Eurasian livestock and many New World crops. Except for barley and flax, the wild ancestors of the Fertile Crescent founder crops had restricted geographic ranges confined to the area between modern Turkey and western Iran, while chickpea was even more narrowly restricted, to southeastern Turkey. Those small geographic ranges, plus the rapid spread of domesticates along Eurasia’s east–west axis, meant that, once a wild plant had been domesticated, it spread so rapidly that further independent domestications of the same or related species were pre-empted. The large Eurasian mammals, however, had such wide geographic ranges (in the case of pigs extending for 13,000 km from Spain to China) that there was ample time for independent domestications at locations west and east of each other. In the New World, even though all the homelands of agriculture lay within only 4000 km of each other, the slowness of crop diffusion along the New World’s north–south axis meant that repeated independent domestications were frequent. So slow was that diffusion that the New World’s main animal domesticates — the llama and guinea pig of the Andes, and the turkey of Mexico — had

not even spread the mere 2,000 km north to Mexico and south to the Andes, respectively, by the time that Europeans arrived in AD 1492.

Primary versus secondary homelands

In several parts of the world, food production arose only upon the arrival of domesticates from the primary homelands, whereupon people proceeded to domesticate some local wild plants or animals that had not been domesticated previously⁹. Clear examples of such ‘secondary’ homelands, in which local domestication was triggered by the arrival of Fertile Crescent crops, were Europe (local domestication of poppies and possibly oats) and Egypt (chufa and sycamore fig).

The recognition of those secondary homelands requires us to reconsider the supposed primary homelands. On the one hand, some of the primary homelands may better be viewed as consisting of multiple homelands in which distinct systems of food production arose nearby but independently of each other. This is especially true for the homeland of Andes/Amazonia, which actually comprised primary highland sites in the Andes as well as primary lowland sites scattered from Panama through the Amazon Basin to the Pacific coast of Ecuador and Peru^{55,56}. Similarly, the Mesoamerican and Fertile Crescent homelands may have consisted of a mixture of highland and lowland sites, while China probably included northern and southern sites in the Yellow River and Yangtze River basins, respectively, as well as coastal lowland and interior upland sites.

On the other hand, some of the nine candidates for primary homelands may actually be secondary homelands in which domestication was triggered by the arrival of domesticates or of farmers from elsewhere. Independent origins of food production seem indisputable for five of the candidates (the Fertile Crescent, China, Mesoamerica, South America and eastern United States), because they were the earliest sites of domestication in their respective parts of the world. But questions have been raised, at least in conversation, regarding the independence of the other four candidates. Especially uncertain is the status of Ethiopia, where it is unknown whether several undoubted local domesticates (teff, coffee, finger millet, chat, noog and ensete) were cultivated before or only after the arrival of Fertile Crescent domesticates, and the New Guinea highlands, where remains of irrigation and drainage systems attest to early agriculture but where the first crops grown remain unidentified and the earliest dates of food production remain disputed. The independence of even the eastern United States has been challenged recently^{42,57}, but the evidence seems compelling that Mexican crops arrived there only by way of southwestern United States and only long after local eastern origins of domestication^{8,58}. Conversely, in southern India the exact dates of arrival of Fertile Crescent domesticates and of earliest cultivation of local domesticates remain uncertain.

Mechanism of the spread of food production

As already noted, the spread of agriculture from its homelands involved in a few cases the acquisition of domesticates by hunter–gatherers outside the homelands, and in more cases the spread of farmers themselves from the homelands. The contributions of these two processes await resolution in many other cases. For example, contrary to what I wrote five years ago¹, the spread of farming in coastal west Mediterranean Europe (in the form of the Cardial and impressed ware cultures) now seems to have involved the rapid transport by sea of a complete package of Neolithic domesticates around 5400 BC by colonizing pioneer farmers⁵⁹. The Yayoi horizon, which marks the arrival of intensive rice agriculture in Japan, and which Japanese scholars until recently preferred to view as an adoption of mainland practices by the indigenous pre-existing Japanese population, now seems increasingly likely on genetic evidence to represent the arrival, population increase and spread of Korean farmers³³.

Why large domestic mammals were mainly Eurasian

Part of the reason why large domestic mammals were mainly Eurasian is simply that Eurasia, being the largest continent and

having escaped the Late-Pleistocene extinctions that eliminated most large mammal species of the Americas and Australia⁶⁰, has the largest number of large wild mammal species. But there is a second part to the answer — a much higher percentage of large mammal species proved domesticable in Eurasia (18%) than in any other continent (Table 9.2 of ref. 1). Especially striking is the contrast of Eurasia with sub-Saharan Africa, where none of the 51 large mammal species was domesticable.

This difference constitutes a problem not in human behaviour, but in animal behaviour and sociobiology — something about African environments selected for one or more of the six mammalian traits that made domestication difficult. We already have some clues, as many of Africa's large mammals are species of antelopes and other open-country mammals whose herds lack the follow-the-leader dominance hierarchies characterizing Eurasian cattle, sheep, goats and horses^{3,61}. To resolve this problem, I suggest attempting to assign one or more of the six traits derailing domestication to each of the non-domesticated large mammal species of Eurasia and Africa, then evaluating the environmental factors behind the evolution of that trait.

History of domestication of particular species

The history of domestication is much better understood for domesticates of western Eurasia than of other parts of world. Taking Zohary & Hopf's³ account of western Eurasian plant domestication as a gold standard, it will be a challenge to workers on other biotas to match that standard. Even for western Eurasia, important unanswered questions abound. To mention only one out of dozens, calculation of molecular divergence times between dogs and wolves suggests that domestication of wolves began around 100,000 years ago^{62,63}, yet the marked morphological differences between wolves and dogs (which should be easily detectable in fossilized skeletons) do not appear until about 11,000 years ago. How can the molecular data and the morphological data be reconciled?

The Future of domestication

Further domestications of plants and animals

We humans today depend for our survival on that tiny fraction of wild species that has been domesticated. Might the rise of molecular biology, genetics and understanding of animal behaviour help feed our growing numbers by increasing that tiny fraction? Modern science has indeed made it technically possible to 'domesticate' species undomesticable in the past, in the sense that we achieve far more draconian control over the captive breeding of endangered California condors (computer-matched for mating to maximize genetic diversity) than the low-tech control that ancient animal breeders exerted over their livestock. But although this 'domestication' is of great interest to conservation biologists, it holds no promise of a condor industry to displace chicken from the supermarkets. What wild species might now be domesticated with profit?

It is instructive to reflect on the meagre additions to our repertoire of domestic species in recent millennia, despite monumental efforts. Of the world's 14 valuable big domestic mammals, the sole addition within the last millennium has been the reindeer, one of the least valuable of the 14. (In contrast, the five most valuable — the sheep, goat, cow, pig and horse — had all been domesticated repeatedly by 4000 BC.) Long-ongoing efforts by modern livestock breeders to domesticate other large wild mammals have resulted either in virtual failure (for example, eland, elk, moose, musk ox and zebra), or else in ranching animals (deer and American bison) that still cannot be herded and that remain of trivial economic value compared to the five most valuable mammals. Instead, all of the mammalian species that have recently become well established as domesticates (for example, arctic fox, chinchilla, hamster, laboratory rat and rabbit) are small mammals dwarfed in usefulness as well as in size by cows and sheep. Similarly, whereas several wild plants were first domesticated only in modern times (for example, blueberries, macadamia nuts, pecans

and strawberries), their value is insignificant compared to that of ancient domesticates such as wheat and rice.

Our best hopes for valuable new domesticates lie in recognizing the specific difficulties that previously derailed domestication of particular valuable wild species, and using modern science to overcome those difficulties. For instance, now that we understand the polygenic control of non-bitterness in acorns, perhaps we could use that knowledge to select for oaks with non-bitter acorns, just as ancient farmers selected for non-bitterness controlled by a single gene in almonds. I am concerned, however, that such attempts may in the long run do us more harm than good. Humanity's greatest risk today is of our growing numbers and aspirations ultimately destroying our society by destroying our environment. Providing undernourished people with more food would be a laudable goal if it were inexorably linked to reducing our numbers, but in the past more food has always resulted in more people. Only when crop and animal breeders take the lead in reducing our numbers and our impacts will they end up by doing us net good.

Further domestication of humans

Some genotypes that used to serve us well as hunter-gatherers now serve us poorly as first-world citizens who forage only in supermarkets — especially metabolically thrifty genotypes that now predispose to type II diabetes, salt-conserving genotypes that predispose to hypertension, and other genotypes predisposing to other cardiovascular diseases and lipid disorders. As formerly spartan populations become westernized ('coca-colonized')⁶⁴, they fall victim to these diseases of the western lifestyle, extreme examples being the 70% incidence of type II diabetes in those Nauru Islanders and Pima Indians lucky enough to survive to the age of 60 (ref. 65). Because diabetes now afflicts south Asians and Pacific Islanders already in their twenties with high morbidity and mortality, there has been detectable natural selection against the predisposing genotypes even within just recent decades. The lower frequency of type II diabetes in Europeans than in non-Europeans matched for diet and lifestyle suggests that natural selection had already reduced European frequencies of those genotypes in previous centuries, as the western lifestyle was developing in Europe. In effect, the unconscious domestication of humans by agriculture that began over 10,000 years ago is still underway.

Even more such gene-frequency changes, also known as illness and deaths, are expected in the near future, as westernization accelerates in the world's two most populous countries, China and India^{66,67}. For example, the incidence of type II diabetes in mainland China, until recently less than 1%, has already tripled in some areas. What lies ahead for China can be projected by considering overseas Chinese populations in Hong Kong, Taiwan, Singapore and Mauritius, where westernization is further advanced and the incidence of type II diabetes is up to 17%. Similarly, the incidence in overseas Indian populations such as that of Fiji gives a foretaste of diabetes' future in India itself.

The resulting projections are that the number of cases of diabetes is expected to increase worldwide by 46% from the year 2000 to 2010, to reach around 220 million in 2010 and around 300 million in 2025. The steepest increase will be in east Asia (including China and India), the projected home of 60% of the world's diabetics in 2010. Similar diet-related disease epidemics are underway in less numerous peoples (from Africans to Aboriginal Australians), involving not just diabetes but also hypertension and other conditions. Thus, these epidemics pose the same dilemma as do efforts to domesticate more wild plant and animal species: how can we ensure that agriculture spreads only happiness, and not suffering as well? □

doi:10.1038/nature01019

1. Diamond, J. *Guns, Germs, and Steel: the Fates of Human Societies* (Norton, New York, 1997).
2. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton Univ. Press, Princeton, 1994).
3. Clutton-Brock, J. *Domesticated Animals From Early Times* (Univ. Texas Press, Austin, 1981).
4. Kiple, K. F. & Ornelas, K. C. *The Cambridge World History of Food* (Cambridge Univ. Press, Cambridge, 2000).
5. Ladizinsky, G. *Plant Evolution under Domestication* (Kluwer, Dordrecht, 1998).

6. Ruhlen, M. *A Guide to the World's Languages* (Stanford Univ. Press, Stanford, 1997).
7. Smartt, J. & Simmonds, N. W. *Evolution of Crop Plants* 2nd edn (Longman, Harlow, 1995).
8. Smith, B. D. *The Emergence of Agriculture* (Scientific American Library, New York, 1998).
9. Zohary, D. & Hopf, M. *Domestication of Plants in the Old World* 3rd edn (Oxford Univ. Press, Oxford, 2000).
10. Cohen, M. N. & Armelagos, G. J. *Paleopathology at the Origins of Agriculture* (Academic, Orlando, 1984).
11. Diamond, J. *The Third Chimpanzee: the Evolution and Future of the Human Animal* (HarperCollins, New York, 1992).
12. Darwin, C. *The Origin of Species by Means of Natural Selection* (Murray, London, 1859).
13. Rindos, D. *The Origins of Agriculture: An Evolutionary Perspective* (Academic, San Diego, 1984).
14. Flannery, K. V. in *The Domestication of Plants and Animals* (eds Ucko, P. J. & Dimbleby, G. W.) 73–100 (Duckworth, London, 1969).
15. Binford, L. F. in *New Perspectives in Archaeology* (eds Binford, S. R. & Binford, L. R.) 313–341 (Aldine, Chicago, 1968).
16. Stiner, M. C., Munro, N. D., & Surovell, T. A. The tortoise and the hare: small-game use, the broad-spectrum revolution, and paleolithic demography. *Curr. Anthropol.* **41**, 39–73 (2000).
17. Hillman, G. C., & Davies, M. S. Measured domestication rates in wild wheats and barley under primitive cultivation, and their archaeological implications. *J. World Prehist.* **4**, 157–222 (1990).
18. Smith, B. D. Low-level food production. *J. Archaeol. Res.* **9**, 1–43 (2001).
19. Smith, B. D. Documenting plant domestication: the consilience of biological and archaeological approaches. *Proc. Natl Acad. Sci. USA* **98**, 1324–1326 (2001).
20. Marchetti, M. & Nevitt, G. Effect of hatchery rearing on brain structures of rainbow trout, *Oncorhynchus mykiss*. *Env. Biol. Fishes* (in the press).
21. Diamond, J. Competition for brain space. *Nature* **382**, 756–757 (1996).
22. Clutton-Brock, J. *Horse Power* (Harvard Univ. Press, Cambridge, 1992).
23. Lev-Yadun, S., Gopher, A., & Abbo, S. The cradle of agriculture. *Science* **288**, 1602–1603 (2000).
24. Marshall, F. in *The Origins and Development of African Livestock* (eds Blench, B. M. & MacDonald, K. C.) 191–221 (UCL Press, London, 2000).
25. McNeill, W. *Plagues and Peoples* (Doubleday, Garden City, 1976).
26. Crosby, A. *The Columbian Exchange: Biological Consequences of 1492* (Greenwood, Westport, 1972).
27. Crosby, A. *Ecological Imperialism: The Biological Expansion of Europe, 900–1900* (Cambridge Univ. Press, Cambridge, 1986).
28. Ramenofsky, A. *Vectors of Death* (Univ. New Mexico Press, Albuquerque, 1987).
29. Ehret, C. *An African Classical Age* (Univ. Press of Virginia, Charlottesville, 1998).
30. Bellwood, P. *Prehistory of the Indo-Malaysian Archipelago* revised edn (Univ. Hawaii Press, Honolulu, 1997).
31. Ammerman, A. J. & Cavalli-Sforza, L. L. *Neolithic Transition and the Genetics of Populations in Europe* (Princeton Univ. Press, Princeton, 1984).
32. Renfrew, C. *Archaeology and Language: The Puzzle of Indo-European Origins* (Cambridge Univ. Press, New York, 1987).
33. Hudson, M. *Ruins of Identity* (Univ. Hawaii Press, Honolulu, 1999).
34. Vogel, F. & Chakravarti, M. R. ABO blood groups and smallpox in a rural population of West Bengal and Bihar (India). *Hum. Genet.* **3**, 166–180 (1966).
35. Diamond, J. & Rotter, J. I. in *The Genetic Basis of Common Diseases* (eds King, R. A., Rotter, J. I. & Motulsky, A. G.) 50–64 (Oxford Univ. Press, New York, 2002).
36. Eaton, S. B., Shostak, M. & Konner, M. *The Paleolithic Prescription* (Harper and Row, New York, 1988).
37. Bonsall, C. *The Mesolithic in Europe* (John Donald, Edinburgh, 1990).
38. Lourandos, C. *Continuum of Hunter-Gathers: New Perspectives in Australian Prehistory* (Cambridge Univ. Press, Cambridge, 1997).
39. Arnold, J. E. *The Origins of a Pacific Coast Chiefdom: The Chumash of the Channel Islands* (Univ. Utah Press, Salt Lake City, 2001).
40. Klein, R. G. *The Human Career: Human Biological and Cultural Origins* 2nd edn (Univ. Chicago Press, Chicago, 1999).
41. Richerson, P. F., Boyd, R. & Bettinger, R. L. Was agriculture impossible during the Pleistocene but mandatory during the Holocene? A climate change hypothesis. *Am. Antiquity* **66**, 387–412 (2001).
42. Sanjurjo, O. I., Piperno, D. R., Andres, T. C. & Wessel-Beaver, L. Phylogenetic relationships among domesticated and wild species of *Cucurbita* (Cucurbitaceae) inferred from a mitochondrial gene: implications for crop plant evolution and areas of origin. *Proc. Natl Acad. Sci. USA* **99**, 535–540 (2002).
43. Zohary, D. in *The Origins and Spread of Agriculture and Pastoralism in Eurasia* (ed. Harris, D. R.) 142–158 (UCL Press, London, 1996).
44. Zohary, D. Monophyletic vs. polyphyletic origin of the crops on which agriculture was founded in the Near East. *Genet. Resources Crop Evol.* **46**, 133–142 (1999).
45. Heun, M., Schäfer-Pregl, R., Klawan, D. et al. Site of Einkorn wheat domestication identified by DNA fingerprinting. *Science* **278**, 1312–1314 (1997).
46. Loftus, R. T. et al. Evidence for two independent domestications of cattle. *Proc. Natl Acad. Sci. USA* **91**, 2757–2761 (1994).
47. Lau, C. H. et al. Genetic diversity of Asian water buffalo (*Bubalus bubalis*): mitochondrial DNA D-loop and cytochrome b sequence variation. *Anim. Genet.* **29**, 253–264 (1998).
48. Vilà, C. et al. Widespread origins of domestic horse lineages. *Science* **291**, 474–477 (2001).
49. MacHugh, D. E. & Bradley, D. G. Livestock genetic origins: goats buck the trend. *Proc. Natl Acad. Sci. USA* **98**, 5382–5384 (2001).
50. Troy, C. S. et al. Genetic evidence for Near Eastern origins of European cattle. *Nature* **410**, 1088–1091 (2001).
51. Luikart, G. et al. Multiple maternal origins and weak phylogeographic structure in domestic goats. *Proc. Natl Acad. Sci. USA* **98**, 5927–5932 (2001).
52. Giuffra, E. et al. The origin of the domestic pig: independent domestication and subsequent introgression. *Genetics* **154**, 1785–1791 (2000).
53. Hanotte, O. et al. African pastoralism: genetic imprints of origins and migrations. *Science* **296**, 336–339 (2002).
54. Hiendler, S. et al. Molecular analysis of wild and domestic sheep questions current nomenclature and provides evidence for domestication from two different subspecies. *Proc. R. Soc. Lond. B* **269**, 893–904 (2002).
55. Piperno, D. R. & Pearsall, D. M. *The Origins of Agriculture in the Lowland Neotropics* (Academic, San Diego, 1998).
56. Piperno, D. R., Rinere, A. J., Holst, I. & Hansell, P. Starch grains reveal early root crop horticulture in the Panamanian tropical forest. *Nature* **407**, 894–897 (2000).
57. Lentz, D. et al. Prehistoric sunflower (*Helianthus annuus* L.) domestication in Mexico. *Econ. Bot.* **55**, 370–376 (2001).
58. Smith, B. D. *Rivers of Change: Essays on Early Agriculture in Eastern North America* (Smithsonian Institution Press, Washington DC, 1992).
59. Zilhão, J. Radiocarbon evidence for maritime pioneer colonization at the origins of farming in West Mediterranean Europe. *Proc. Natl Acad. Sci. USA* **98**, 14180–14185 (2001).
60. Martin, P. S., & Klein, R. G. *Quaternary Extinctions* (Univ. Arizona Press, Tucson, 1984).
61. Spinnage, C. *The Natural History of Antelopes* (Fax on File, New York, 1986).
62. Vilà, C. et al. Multiple and ancient origins of the domestic dog. *Science* **276**, 1687–1689 (1997).
63. Vilà, C., Maldonado, J. E. & Wayne, R. K. Phylogenetic relationships, evolution, and genetic diversity of the domestic dog. *J. Hered.* **90**, 71–77 (1999).
64. Zimmet, P. Globalization, coca-colonization and the chronic disease epidemic: can the Domsday scenario be averted? *J. Intern. Med.* **247**, 301–310 (2001).
65. Dowse, G. K., Zimmet, P. Z., Finch, C. F. & Collins, V. R. Decline in incidence of epidemic glucose intolerance in Nauruans: implications for the “thrifty genotype”. *Am. J. Epidemiol.* **133**, 1093–1104 (1991).
66. Amos, A., McCarthy, D. & Zimmet, P. The rising global burden of diabetes and its complications: estimates and projections to the year 2010. *Diabetic Med.* **14**, S1–S5 (1997).
67. Zimmet, P., Alberti, K. G. M. M. & Shaw, J. Global and societal implications of the diabetes epidemic. *Nature* **414**, 782–787 (2001).

Acknowledgements

It is a pleasure to acknowledge my debt to P. Bellwood, A. Ehrlich, K. Flannery, I. Hodder and B. Smith for valuable suggestions.

Contacts

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Senior European Sales Manager:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)
Andy Douglas (4975)
Frank Phelan (4944)
Amelie Pequignot (4974)
Laura Pearson (4977)

Netherlands/ Italy/ Iberia:

Evelina Rubio Hakansson (4973)
Scandinavia: Silje Opstrup (4994)

Production Manager: Billie Franklin

To send materials use London
address above.
Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

International

Advertising Coordinator:

Hind Berrada (4935)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Offices

France/ Belgium:

Christine Niox-Chateau
Tel +33 (0) 1 43 20 16 51
Fax +33 (0) 1 43 20 51 52
e-mail: c.nioxchateau@nature.com

Germany/ Austria/ Switzerland:

Patrick Phelan, Odo Wulffen
Tel +49 89 54 90 57 11/-2
Fax +49 89 54 90 57 20
e-mail: p.phelan@nature.com
o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager: Peyton Mason

US Advertising Coordinator:

Ashly de Leon

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Harakata-machi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746
e-mail: k.johnson@naturejpn.com

Asia-Pacific Advertising Manager:

Kevyn Johnson

naturejobs

Lessons in career planning

Faced with a glut of science PhDs and a shortage of teachers in both elementary and high schools, the US National Research Council (NRC) has taken the logical step of using one problem to solve the other. In a report released on 30 July, the council advocates the creation of a fellowship programme to draw PhDs into teaching. As well as solving the present supply-and-demand problems in both areas, the report says, the programme would improve the quality of science and maths education across the United States.

The NRC proposes a two-year programme that would pay new PhDs a stipend of \$35,000 a year, which is competitive with the funding received by some research postdocs. It would also include the training needed to qualify as a teacher.

But the report doesn't fully answer one crucial question: do recent PhDs want to make the transition from research to teaching? Although the NRC anticipates "a high level of interest" for the programme in the short term, it fails to explore the long-term ramifications for the PhDs' career paths.

As proposed, the programme could very well hamper PhDs should they want to return to research. New PhDs who do not publish much early in their careers tend to have a difficult time securing research posts later on. Such problems are already faced by PhDs who consider postdoc positions that emphasize teaching at universities (see *Naturejobs* 5; 28 February 2002).

What, then, is a sensible alternative? Perhaps a more flexible programme that splits a new PhD's time between a research institution and a high school. Such an arrangement would allow the PhD to keep a foot in the scientific world that, ostensibly, should enhance their teaching ability. And, in turn, it would keep their career options open for when their two-year term is up.

Paul Smaglik

Naturejobs editor



Contents

SPECIAL REPORT

Attractive forces in Asia p4

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

SCIENTIFIC ANNOUNCEMENTS

SCIENTIFIC EVENTS



Pacific Rim countries are vying to recruit scientists, but they face some stiff obstacles, says Robert Triendl.



Capital Biochip's products have won it respect in the West.



Both Yoshiaki Ito (above) and Cheng Jing (right, below) are benefiting from a fresh funding impetus for life sciences in Asia.

Yoshiaki Ito, one of Japan's leading cancer researchers, did something very unusual when he retired from his post at Kyoto University in March. Standard practice would have seen him retire to a comfortable, well-paid teaching job at a private establishment. But Ito, together with seven members of his research team, accepted an offer to establish a research laboratory at the National University of Singapore.

This unusual move emphasizes both Singapore's bold vision for life sciences and the increasingly active jobs market in the Pacific Rim. Singapore's expansion in scientific research is being fuelled by a combination of the right people and an aggressive funding system, says Sydney Brenner, a molecular biologist who advises the Singapore government on its life-sciences investments. And as other Asian and Pacific Rim countries concentrate on establishing a scientific infrastructure, they too will make a more concerted effort to attract — or bring back — leading scientists to the region, he says.

As well as being a professor at the Salk Institute for Biological Studies in San Diego, Brenner heads a laboratory at Singapore's Institute of Molecular and Cell Biology. He is one of several prominent Western scientists who have found their way to Asia, and was instrumental in bringing research on the pufferfish genome to Singapore.

Over the past few years, Asian countries such as China and Singapore have launched initiatives — often with generous inducements and the promise of significant financial support — to attract back native scientists who left to pursue their careers abroad. For some, the prospect of a special relationship with funding agencies and patrons, and the promise of access to the money, people and independence necessary to make truly revolutionary contributions, is proving to be irresistible.

A US citizen born in Hong Kong, Edison Liu is a former director in the

Division of Clinical Sciences at the US National Cancer Institute who now heads the Genome Institute of Singapore. He says he was drawn to Singapore by the prospect of realizing a new vision of genome research — and by the challenge of building an entire genomics centre from scratch, backed by a government willing to spend large amounts on life-sciences research. "Many people have been talking about the prospect of linking genome research with cell biology, but there are hardly any places around that have actually attempted to realize that synthesis," Liu says. He believes the institute in Singapore is able to fulfil this dream.

Prominent scientists returning to China have also received both political backing and significant funding. *Fortune Magazine* has named Capital Biochip, a state-funded venture in Beijing that

Closed shops opening up?

Japanese universities have long been "closed shops", according to cultural historian Ivan Hall, a long-term resident of Japan. A few years ago, his book *Cartels of the Mind*, about Japan's inward-looking universities and intelligentsia, provoked angry responses from the government — and an awkward silence from academics.

Several factors contribute to keeping down the number of foreign staff — with the exception of English teachers — at national universities. Posts available to foreign scientists are few and salaries remain quite low. And not many

foreigners are willing to take the challenge of learning Japanese.

But there are signs that opportunities are increasing. University and research institutes offer a steadily growing number of positions for short- and long-stay visitors, and some large government-funded research centres have hired foreign scientists as lab or division directors.

An upcoming university reform that will provide more autonomy to national universities, and that aims to infuse more competition into Japanese academia, could help to open up the system further. **R.T.**

CAPITAL BIOCHIP

NATL UNIV. SINGAPORE

CAPITAL BIOCHIP



Top jobs

The Brazilian Carlos Ghosn is seen as a corporate hero in Japan for his feat in rescuing the ailing Nissan motor company. But many Japanese companies are reluctant to appoint a foreigner to a senior management position — even at their overseas subsidiaries.

George Stephanopoulos, chief technology officer at Mitsubishi Chemical, Japan's leading chemical company, is a rare exception. A professor of chemical engineering at the Massachusetts Institute of Technology for the past three years, Stephanopoulos has been responsible for running Mitsubishi's research

and development (R&D) division.

Some observers say that Mitsubishi, reputedly a conservative company where change doesn't come easily, had hired him mainly to clean up, and downsize, R&D operations. But Stephanopoulos has already instilled a number of fresh ideas into its R&D operations, such as a broad cooperation agreement with universities in Japan and the United States. And, under his leadership, the company has started to reconsider its life-sciences strategy.

Stephanopoulos calls his position there the "most exciting experience in my life". R.T.

Attractive prospects: equipment such as the Japanese Earth Simulator (left) and initiatives such as the project to sequence the pufferfish genome (the team is shown on the right) are drawing scientists to the Pacific Rim.

IMCB

develops lab-on-a-chip systems, the "coolest international biotech company", which seems to testify that the strategy is working.

Set up by Cheng Jing, formerly a research scientist with the San Diego start-up company Nanogen, Capital Biochip is one of several state-backed life-sciences ventures to have received significant funding from the Chinese government and to have persuaded prominent Chinese scientists to return from overseas.

INTERNATIONAL CENTRES

In Japan, the economic and technological powerhouse of the Pacific Rim, the newly opened Center for Developmental Biology in Kobe, part of the Institute of Physical and Chemical Research (RIKEN), has hired several Japanese scientists formerly affiliated with US universities as directors of laboratories or divisions. The new centre provides excellent funding and the chance to build an organization from scratch.

Companies in Japan are increasingly looking to foreign scientists, or Japanese scientists who have made their career abroad, to aid interaction with universities or start-up companies abroad and to strengthen their ties with academia (see 'Top jobs', right).

In some cases it is large-scale facilities or unique equipment that bring scientists to Japan. For example, Earth scientists and computer specialists around the globe are eager to access the Japanese Earth Simulator, the world's most powerful computer system dedicated to the geosciences (see *Naturejobs* 7; 30 May 2002). With the new system now available, filling research positions reserved for foreign scientists at the Earth Simulator Research Center or the adjacent Frontier Research System for Global Change should not be difficult.

Similarly, the nuclear magnetic resonance (NMR) facility at RIKEN's Genomic Sciences Center in Yokohama is already drawing structural genomics scientists to Japan. One of them is Peter Guntert, a specialist in NMR data analysis who used to work at the Swiss Federal Institute of Technology in Zurich. He now heads a small laboratory at RIKEN. With some 40 state-of-the-art NMR machines, there could hardly be a better place to test his software.

But beyond a relatively small number of these highly visible and well-funded research efforts, it seems that many Japanese research organizations turn away foreign scientists rather than attracting them (see

'Closed shops opening up?', opposite). The notoriously inward-looking scientific establishment even repels some of the brightest home-grown talent.

When Shuji Nakamura, the inventor of the blue laser diode, angrily left Japan for the University of California, Santa Barbara, two years ago, some warned of a new brain drain towards the United States. But, although some of the most brilliant young Japanese scientists, many of them female, find no choice other than to go abroad, the number pursuing their careers in Europe or the United States has hardly changed in the past few years. In the 1950s and 1960s, a PhD or postdoc at a US university was deemed vital to a solid scientific career. But in today's increasingly tight academic market, there is a greater temptation to stay at home rather than lose out on the contacts necessary to secure a rare tenure-track position at the prestigious national universities.

On the other hand, as Ito's case illustrates, prominent scientists in Japan can be tempted to move to locations that offer them the best conditions. Some in Japan dismiss Ito's relocation to Singapore as a one-off case. But nobody disputes that more dynamic labour markets for research professionals will benefit countries across the region — including Japan. And Ito's move may be a step in that direction.

Robert Triendl is a freelance writer based in Tokyo.



Working abroad: Sydney Brenner (above) and George Stephanopoulos (below) have made the transition from West to East.



IMCB

MITSUBISHI CHEMICAL